



LEE JUSSIM

SOCIAL PERCEPTION
and SOCIAL REALITY

Why Accuracy Dominates Bias and Self-Fulfilling Prophecy

OXFORD

SOCIAL PERCEPTION AND SOCIAL REALITY

This page intentionally left blank

Social Perception and Social Reality

WHY ACCURACY DOMINATES BIAS AND
SELF-FULFILLING PROPHECY

Lee Jussim

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

*Oxford University Press, Inc., publishes works that further
Oxford University's objective of excellence
in research, scholarship, and education.*

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi Kuala Lumpur Madrid Melbourne
Mexico City Nairobi New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece Guatemala Hungary Italy
Japan Poland Portugal Singapore South Korea Switzerland Thailand Turkey Ukraine Vietnam

Copyright © 2012 by Oxford University Press, Inc.

Published by Oxford University Press, Inc.
198 Madison Avenue, New York, New York 10016

www.oup.com
Oxford is a registered trademark of Oxford University Press, Inc.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or
transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise,
without the prior permission of Oxford University Press

Library of Congress Cataloging-in-Publication Data

Jussim, Lee J.
Social perception and social reality : why accuracy dominates bias and self-fulfilling prophecy/Lee Jussim.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-19-536660-0 (hardcover)
1. Social perception. 2. Stereotypes (Social psychology) I. Title.
BF323.S63J87 2012
302'.12—dc23 2011020796

9 8 7 6 5 4 3 2 1

Printed in the United States of America on acid-free paper

This book is dedicated to my family: my wife, Lisa Baum, who has not merely been supportive, but with whom conversations about multiculturalism in her practice influenced my thinking about stereotypes; my daughter Rachel, whose righteous indignation at social scientists' resistance to scientific evidence right in front of their noses has been an ongoing inspiration for me; my daughter Kayla, whose spunk and resilience in the face of extraordinary difficulties is its own constant source of inspiration for me; and to my son Josh, for giving me numerous opportunities to discover that his very multi-cultural set of friends are not remotely threatened by talking about ethnic and cultural differences between groups and, in fact, generally enjoy such discussions.

This page intentionally left blank

Acknowledgments

MANY OF THE ideas in the book are so outside the mainstream of “normal” social science claims, that it is perhaps more obvious than usual that any errors, misinterpretations, misrepresentations, and the like are entirely my own. For example, this book spends 6 chapters arguing that the effects of expectations are typically overstated and, rather than being powerful and pervasive, are weak, fragile, and fleeting. It spends 3 chapters arguing that there is far more evidence of accuracy than most social psychologists acknowledge. It spends 5 chapters arguing that stereotypes are typically quite accurate, typically used in a manner that is reasonable and more or less rational, and that it is social psychological perspectives emphasizing stereotype inaccuracy that are exaggerated, unjustified, and irrationally resistant to change. One can find vanishingly few psychologists of any stripes presenting such claims, so that one can be assured that any errors in making or justifying them are entirely my own.

Although they bear no responsibility whatsoever for any of the claims that appear throughout the book, several people have been immensely helpful in preventing me from making claims that even I would consider unjustified. I am deeply grateful for critiques and commentary received on one or more chapters from David Funder, Bill von Hippel, David Kenny, Joachim Krueger, Clark McCauley, Richard Nisbett, Charles Stangor, and Bill Swann. I (and this book) have also benefitted from correspondences with Alice Eagly, Sam Gosling, Judy Hall, C. Neil Macrae, Beth Morling, Michael Norton, Stephen Raudenbush, Carey Ryan, and Sam Sommers.

This book took so long to write that, while still in progress, I extracted core ideas from and have published them elsewhere (*Advances in Experimental Social Psychology*, *Personality and Social Psychology Review*, and the *Handbook of Stereotypes, Prejudice, and Discrimination*). The critical comments provided by anonymous reviews, and by Bill Ickes, Todd Nelson, and Mark Zanna regarding those published pieces were then fed back to improve the chapters in this book. I am very grateful for those thoughtful and cogent comments.

This book is, in part, a scholarly and intellectual (rather than political) polemic. The dictionary definition of polemic is: a controversial argument, as one against some opinion, doctrine, etc. This book definitely argues against the pervasive view among social cognition scholars that human social judgment is dominated by bias; and it even argues against the pervasive doctrine in psychology and other social sciences that stereotypes are inaccurate

and irrational. As such, I must also thank all those scholars who have promoted and advocated the view of human social cognition as deeply flawed and steeped in error and bias—without them, I would have had no reason to write this book.

Of course, the book does not mostly “argue against” anything—it mostly argues for a view of human social thinking as generally nicely in touch with social reality and, generally, subject to biases that, though real, are readily reduced or eliminated. Although my own original research constitutes a tiny fraction of that reported in this book, much of that research was conducted with Jacque Eccles and Stephanie Madon, collaborations for which I am deeply grateful. Last, I must thank the Hillsborough Diner in NJ, where I wrote most of this book, for allowing me to sit and write for many hours at a time, despite only spending a few bucks at a pop for breakfast or lunch.

Contents

PART ONE | INTRODUCTION: THIS BOOK, BASIC IDEAS, AND THE EARLY RESEARCH

- 1 *Introduction: How Might Social Beliefs Relate to Social Reality?* 3
- 2 *Social Reality Is Not Always What It Appears To Be: The Scientific Roots of Research on Interpersonal Expectancies* 13
- 3 *The Once Raging and Still Smoldering Pygmalion Controversy* 30

PART TWO | THE AWESOME POWER OF EXPECTATIONS TO CREATE REALITY AND DISTORT PERCEPTIONS

- 4 *The Extraordinary Power of Self-Fulfilling Prophecies* 49
- 5 *The Extraordinary Power of Expectancies to Bias Perception, Memory, and Information-Seeking* 64

PART THREE | THE LESS THAN AWESOME POWER OF EXPECTATIONS TO CREATE REALITY AND DISTORT PERCEPTIONS

- 6 *The Less Than Extraordinary Power of Self-Fulfilling Prophecies: Considerations Based on Common Sense, Daily Life, and a Critical Evaluation of the Early Classic Experiments* 83
- 7 *You Better Change Your Expectations Because I Will Not Change (Much) to Fit Your Expectations: Self-Verification as a Limit to Self-Fulfilling Prophecies* 100
- 8 *The Less Than Awesome Power of Expectations to Distort Information-Seeking* 112
- 9 *The Less Than Awesome Power of Expectations to Bias Perception, Memory, and Judgment* 122

PART FOUR | ACCURACY: CONTROVERSIES, CRITICISMS, CRITERIA, COMPONENTS, AND COGNITIVE PROCESSES

- 10 *Accuracy: Historical, Political, and Conceptual Objections* 145
- 11 *Accuracy: Criteria* 170
- 12 *Accuracy: Components and Processes* 194

PART FIVE | THE QUEST FOR THE POWERFUL SELF-FULFILLING PROPHECY

- 13 *Teacher Expectations: Accuracy and the Quest for the Powerful Self-Fulfilling Prophecy* 219
- 14 *Do Self-Fulfilling Prophecies Accumulate or Dissipate?* 248

PART SIX | STEREOTYPES

- 15 *On the Pervasiveness and Logical Incoherence of Defining Stereotypes as Inaccurate* 269
- 16 *What Constitutes Evidence of Stereotype Accuracy?* 307
- 17 *Pervasive Stereotype Accuracy* 323
- 18 *Stereotypes and Person Perception: Can Judging Individuals on the Basis of Stereotypes Increase Accuracy?* 360
- 19 *Stereotypes Have Been Stereotyped!* 389

PART SEVEN | CONCLUSION

- 20 *Important, Interesting, and Controversial Work on Accuracy, Bias, and Self-Fulfilling Prophecies That Did Not Fit Elsewhere* 407
- 21 *The 90% Full Glass Contests the Bias for Bias* 421

REFERENCES 433

INDEX 455

1 Introduction

THIS BOOK, BASIC IDEAS, AND
THE EARLY RESEARCH

This page intentionally left blank

1 Introduction

HOW MIGHT SOCIAL BELIEFS RELATE TO SOCIAL REALITY?

IS THE GLASS half full or half empty? As everyone knows, the optimist says “half full”; the pessimist says “half empty.” I both love and hate this parable. I love it because it is a terrific metaphor for social perception (how people perceive, judge, evaluate, and understand other people). Are the optimist and pessimist seeing the same glass or a different glass? The way they describe the glass is quite different. Furthermore, their emotional reactions to the “fullness” of the glass are probably quite different. Optimist: “Lord, it’s great to have a half-full glass!” Pessimist: “I can’t believe all I can get is a half-empty glass.”

However, if one looks underneath their tone and their emotions, they are seeing the *exact same objective glass*. The parable is NOT “The glass is half full, but the optimist sees it as 90% full and the pessimist sees it as 10% full.” If that were the case, their respective demeanors would be influencing not merely their respective reactions to the fullness of the glass, but their perceptions of the objective degree of fullness.

Is there a deep and true message here? If so, that message would seem to be that, although our predispositions, demeanors, and expectations can sometimes influence our reactions to events in the world, they do not have much influence on how we perceive the objective characteristics of the events themselves. In this sense, then, the parable is in sharp conflict with many social science and social psychological researchers, who do indeed often claim that our beliefs and expectations powerfully influence and distort our perceptions of objective social reality. The first several chapters of this book document the extraordinary extent to which social scientists have emphasized the power of beliefs to alter not only our perceptions of social reality but also that reality itself.

Which gets me quickly to why I hate this parable. When faced with any sort of intellectual controversy, the easy way to make oneself appear reasonable is to conclude that “there is some

truth to both sides.” Sometimes, this is clearly reasonable and justified by the data. But not always. Lots of conclusions, once believed to be true, turn out to be 100% wrong. Consider, for example, the medieval “spontaneous generation of life” hypothesis, or the idea that the sun revolves around the earth. These ideas were not partially true, half true, or true under some conditions. They were wrong—100% wrong (well, maybe not 100%, in the case of spontaneous generation; credible scientific theories presume that, at some point in the distant past, life did emerge from nonlife. Something that is false nearly all of the time, however, does not undermine my argument at all).

The bottom line is that conclusions should be reached on the basis of data. And when social scientific claims conflict with one another, the extent to which each is true should be determined by data. And the data rarely indicate that one perspective is right half the time and the other is right half the time. Even if both perspectives in some controversy have some truth, one conclusion is typically justified more frequently, or under more common conditions, than the other. And knowing that perspective A is true 90% of the time and perspective B 10% of the time provides a very different view of some conclusion or theory than does the conclusion that “both are true sometimes.”

Are People High or Low Wattage?

Not too long ago, I was discussing with a friend and colleague the widespread assumption in social and much of cognitive psychological scholarship that people are largely irrational, and not only out of touch with reality, but also often not even interested in reality. (This is not the time to justify my characterization of the field as having this view [that will occur throughout the rest of the book], but those interested can refer to Jussim, 1991; Jussim et al., 2005; and Jussim & Harber, 2005, for other places where I have documented the prevalence of such a perspective in much of psychology.) Well, I was not really “discussing” it; actually, I was objecting to it.

And (as he often does), my colleague came up with a terrific turn of phrase. In his words, much of psychological theorizing and scholarship characterizes laypeople as “low wattage,” whereas my own emphasis is that they are “high wattage.” “Low wattage” captures the idea that people are not very bright, that they are lacking in energy and are fundamentally lazy, and that, as a result, they are often irrational and reach invalid conclusions. A more familiar analogy to readers steeped in social and cognitive psychology would be the “cognitive miser.” This idea means different things to different people but, fundamentally, means either or both of two things: (1) people are fundamentally lazy, so they do not do any more cognitive work (thinking, judging, attending, evaluating) than they have to and/or (2) the world is so complex that people need to resort to all sorts of simplifications and corner cutting just to get through the day. It is not that people cannot be intelligent, alert, logical, and rational, but the default is low wattage, low energy, low alertness, and low rationality.

Perspectives emphasizing error, bias, and the ways in which social beliefs create social reality have dominated the literature on social cognition (e.g., Fiske, 1998; Jones, 1986; Kahneman & Tversky, 1973; Nisbett & Ross, 1980; Snyder, 1984). These views have created an image of a social perceiver whose misbegotten beliefs and flawed processes construct not only illusions of social reality in the perceiver’s own mind but also actual social reality through

processes such as self-fulfilling prophecies. In this bleak view, the mind becomes primarily a product of cognitive shortcomings and distorted social interactions.

“High wattage” means just the opposite. People are fundamentally engaged in their social worlds, energetic, and motivated to reach valid conclusions about the world. Of course, people may also be motivated by many things besides accuracy, and they can be overwhelmed or distracted by a very complex social world. Nonetheless, the (minority) high wattage view, to which I largely subscribe, is that although people are certainly not perfect and are subject to systematic and irrational biases and errors, in general, they are socially astute and attempt to thoughtfully negotiate the social world.

Just as the low wattage view does not deny that people can be logical, rational, and in touch with reality, the high wattage view does not deny that people can be illogical, irrational, and out of touch with reality. Thus, the low wattage view does not claim that the glass is 100% empty, and the high wattage view does not claim that the glass is 100% full. Both views, therefore, agree that people are capable of logic and rationality and extraordinary accomplishments; and both agree that people are sometimes irrational and subject to all sorts of distortions and biases. The views differ quite a lot, however, in how they characterize normal and prevalent human social thinking and social perception. By relentlessly emphasizing the empty parts of the glass, the low wattage view is plausibly interpretable as suggesting the glass is mostly empty.

In over 25 years of performing original research and reviewing the evidence on relations between social beliefs and social reality, I have reached the conclusion that psychological and social science data—not the claims or the conclusions, but the data itself—inexorably lead to the conclusion that the glass is 90% full. People are not perfect, but they are pretty damn good. And a large part of my inspiration for writing this book has been to expose some of the extraordinary divergences between the conclusions and emphases of so many social scientists (the low wattage conclusions) and the actual data (which, as far as I can tell, typically paints a picture of people as pretty high wattage).

Are People Mostly In or Out of Touch with Reality?

This is a very big question. Lord knows people believe all sorts of weird stuff—astrology, crystal healing, ghosts, and much much more.

On the other hand, *Homo sapiens* is by far the most successful species on the planet. We dominate every continent except Antarctica. We have taken control of huge tracts of territory, and, generally, when we have not, it has been because we have purposely designated such areas to be “protected” in the forms of parks and preserves. Over the last century, our population has been doubling about every 40 years, and the rate of doubling has been increasing for centuries.

Exactly what has made us so successful? This is also a very large question. Clearly, our intelligence is part of it, although other very intelligent species (gorillas, chimpanzees, bonobos, wolves, many species of whales and dolphins) are not thriving anywhere nearly as well. Our social natures—the extraordinary extents to which we cooperate with and depend on one another—probably also contribute to our success. Again, however, there are many social species on earth, and none are nearly as successful as humans.

Some have begun arguing that it is our propensity for culture that has given us such a huge evolutionary advantage (e.g., Baumeister & Bushman, 2007). Culture refers to the knowledge, wisdom, beliefs, practices, and traditions that are handed down from one generation to the next. Now, admittedly, culture sometimes includes bizarre beliefs and rituals (e.g., human sacrifice to the gods). In general, however, if the information transmitted by cultures was out of touch with reality, it is hard to imagine humans becoming as successful as they have become. So, the existence of some bizarre beliefs and practices probably constitutes dramatic and memorable but nonetheless relatively rare exceptions in the grand scheme of things. Not rare in the sense of “not many people believe or do these things.” Undoubtedly, massive numbers of people have believed all sorts of bizarre things. Rare in the sense that, of all the knowledge that is transmitted from one generation to the next, bizarre and dysfunctional “information” is probably in the minority.

This book, however, is not about human rationality or success writ large. Nor is it about culture. These are too big topics, at least for me at this time. Instead, this book is about one aspect of human reasonableness: our beliefs and expectations for other people. Are these expectations generally in touch with or out of touch with reality? Do people’s expectations for other people usually lead them to distorted perceptions and dysfunctional or self-fulfilling interactions? Or do people generally hold their expectations gently, changing them in response to changing social realities? These questions go to the heart of some fundamental issues about the nature of human social and psychological functioning. They are, therefore, the central questions around which this book is organized.

Three Ways Interpersonal Perceptions Relate to Reality

Accuracy/inaccuracy. One way our social beliefs may relate to social reality is that those beliefs may be accurate or inaccurate. The simple version of accuracy/inaccuracy is that your beliefs about Fred, Akbar, or Nakisha, or about Democrats, Swedes, doctors, or Jews may be right or wrong to varying degrees. Accuracy turns out to be a much more complex issue than it seems, and those complexities will be addressed at length in Chapters 10 through 12. For now, however, it is enough to simply point out that one’s beliefs about other individuals or other groups may be right or wrong to varying degrees.

Accuracy in this book does not refer to people subjectively believing that their beliefs are correct. You can be certain the world is going to end tomorrow, or that the Yankees are going to win the World Series, but if neither happens, then you are wrong. Accuracy does refer to beliefs that correspond well with reality, with one exception. Accuracy (in this book) does not refer to beliefs that lead to their own fulfillment.

Self-fulfilling prophecies. Self-fulfilling prophecies occur when a belief does lead to its own fulfillment. If her dad thinks Alisha is a tennis whiz, and provides her with extensive tennis training, and encourages her to practice tennis for hours each day, Alisha may indeed become a tennis whiz, even if she was no more skilled at tennis than anyone else when she started. Self-fulfilling prophecies are addressed in multiple places throughout this book, but it is enough here to point out that I do not consider them to be a type of accuracy.

Bias. Bias can mean many different things in many different contexts (e.g., race or sex bias, self-serving biases, preferences, etc.). I use it in this book quite narrowly to refer to social

beliefs that influence or distort subjective perceptions and judgments. So, for example, if ninth grade teacher Mr. Jones thinks Ahmed is brilliant, a bias occurs if that belief leads Mr. Jones to judge Ahmed's history essay as better than it really is. A bias (for this book) also occurs if, for example, John's belief that Democrats are more liberal than Republicans causes him (if all other things—especially, policy positions—are equal) to judge a particular Democratic candidate to be more liberal than a particular Republican candidate.

"Bias" is usually a pejorative term. However, in this book, the term is itself inherently neutral. Some biases (such as Mr. Jones's) lead to inaccuracy. Others, such as John's beliefs about the liberalness of Democrats, may increase accuracy. Furthermore, bias differs from self-fulfilling prophecy in one very important respect. A bias occurs when a social belief influences the belief holder's *perceptions* regarding another person, *not* when it influences the other person's actual behavior (that is a self-fulfilling prophecy). Thus, bias involves beliefs influencing perceptions; self-fulfilling prophecy involves beliefs creating an actual reality. These, too, are deep and complex issues and will be dealt with at length in various places throughout this book. For now, it is sufficient to simply define bias as referring to social beliefs that influence or distort ("bias") the perceptions and judgments of the belief holder.

Not mutually exclusive. Accuracy, self-fulfilling prophecy, and bias are not mutually exclusive. Any one, two, or all three can occur simultaneously. For example, the boss might have a pretty accurate view of most of her employees. Pretty accurate, but not perfectly accurate. When she overestimates someone, that might slightly bias her evaluations. These might be important because that employee may receive a larger raise than justified by his record. Furthermore, if he subsequently lives up to the evaluation implied by the raise—that is, his performance actually improves after receiving the raise—then the boss's original belief is also self-fulfilling to some degree. These issues, too, are complex and will be dealt with throughout this book. For now, it is sufficient to point out that accuracy, self-fulfilling prophecy, and bias can occur in any social context in any combination.

What Is Included in This Book

This book is divided into six major sections. The introductory section has two chapters (in addition to this one). Chapter 2 reviews and critically evaluates some of the earliest work on how social beliefs bias judgments, including work on stereotypes, the "New Look" in perception of the 1940s and 1950s (which, for the first time, took seriously the possibility that perception involved distortion and motivation and was not just reception of external stimuli), and some other early and dramatic research. Chapter 3 focuses on the catalyst and springboard for much of the modern scientific interest in relations between social beliefs and social reality: Rosenthal and Jacobson's (1968a,b) dramatic and controversial study of the self-fulfilling effects of teachers' expectations, its aftermath, and, as much as possible, resolutions to the controversies it generated.

Following quickly on the heels of the early teacher expectation research, social psychology fell in love with expectancies. Social psychologists saw self-fulfilling prophecies and expectancy-confirming biases everywhere. The second section, on The Awesome Power of Expectations to Create Reality and Distort Perceptions, attempts to capture and convey

some of the extraordinary enthusiasm for expectancy effects that characterized the field of social psychology in the 1970s and 1980s.

Much of that enthusiasm was, in my view, misplaced—an academic version of stock market “irrational exuberance” (the term coined by former Federal Reserve chair Alan Greenspan to characterize the excessively high stock prices and returns of the late 1990s, just before the crash of 2000). Especially since this type of view still appears fairly frequently in the literature (see Chapter 6), I spend four chapters (The Less Than Awesome Power of Expectations to Create Reality and Distort Perceptions) explaining why neither the original studies nor the subsequent body of research supports strong claims about the power of expectancies.

One reason social beliefs do not typically have powerful or pervasive effects on social reality is that those beliefs are often moderately or even highly accurate. Unfortunately, however, accuracy was long a stigmatized area of research in social psychology, and many a myth grew up around alleged complexities and difficulties in performing accuracy research. Some of those myths contained kernels of truth—for example, assessing accuracy is both more complicated and difficult than it seems at first glance. However, many research areas involve complexities and difficulties, and those involving accuracy are not inordinately worse than those characterizing many other areas of psychological research. Unpacking all this—explaining why accuracy research was stigmatized; identifying the complaints and criticisms often leveled at accuracy research; identifying ways in which the criticisms are true, but also how they can be addressed; identifying the greatly overstated conclusions implying that accuracy research is either not viable or not worth it; and then presenting solutions to the bona fide issues and complexities—takes three chapters (all in the Accuracy section).

By this point, the groundwork has been laid down. The early research has been reviewed, psychology’s early infatuation with expectancy effects has been explored and debunked (or, at least, contested and deconstructed), and the scientific foundations for accuracy research have been established. The next two chapters, then, focus on two issues at the opposite ends of the expectancy spectrum: accuracy and the quest for the powerful self-fulfilling prophecy. Why are teacher expectation effects typically quite limited? There are lots of reasons, but one of them is accuracy. In addition, however, once it became well-established that self-fulfilling prophecy effects were typically much more modest than once thought, the issue of whether they were *ever* powerful became an interesting and important one. Thus, the issues of accuracy and attempts to identify if and when self-fulfilling prophecy effects are ever powerful are reviewed in the section titled The Quest for the Powerful Self-Fulfilling Prophecy.

In a book filled with thorny issues and intellectual/scientific controversies, I saved the best (or worst) for last: stereotypes. Indeed, when I first set out to write this book, I planned on only a single chapter on stereotypes; instead, I ended up with five. This turned out to be necessary, however, because nearly every aspect of stereotypes, including merely defining them, is fraught with cultural and academic myths, logical pitfalls, and bona fide complexities. For example, everyone knows that stereotypes are, by definition, inaccurate, right? Hmmm, well, I do not know that at all. It takes me a whole chapter (Chapter 15) to explain why. Even if they are not by definition inaccurate, aren’t they generally inaccurate in real life?

This is what scientists call “an empirical question”—it requires data to answer. How to obtain such data and analyze it to answer this question is described in Chapter 16, and the data itself is presented in Chapter 17. In many social psychological circles, the idea that stereotypes produce powerful biases in judging individuals is viewed as well-established. Again, however, the extent and power of stereotypes to bias how we judge individuals is an empirical question—one that is thoroughly addressed in Chapter 18. And Chapter 19 addresses the broad implications of all this evidence of stereotype accuracy, rationality, and reasonableness for the real world and the world of psychological theory.

What Is Not in This Book

The subtitle of this book, “Why Accuracy Dominates Bias and Self-Fulfilling Prophecy,” is admittedly ambiguous and could mean lots of different things to different people. Here, therefore, I indicate what is not covered. In general, I do not cover beliefs that are not about specific other people or groups, political beliefs, moral beliefs, or beliefs about which there is no objective reality.

People hold all sorts of beliefs that are odd, invalid, interesting, etc. This book is not about supernatural beliefs, conspiracy theories, out-of-body experiences, and the like. Instead, it is about interpersonal beliefs—people’s beliefs about other people, not about their beliefs regarding things supernatural or generally bizarre.

I suppose political beliefs and ideologies might be considered social to some degree, but they are not what I mean when I use the term “social beliefs.” Whether society should be constructed to ensure that all people have equal rights, equal opportunities, or equal incomes are very interesting topics, but they are not ones that this book addresses. These are moral or philosophical issues, and, for these types of issues, there rarely is an objective social reality; as a result, issues of accuracy, inaccuracy, and self-fulfilling prophecy largely disappear.

Why so many Americans believed that Iraq was involved in the 9/11 attacks is also very interesting and important, but it is also beyond the scope of this book. This book is about people’s beliefs about the characteristics of their friends, acquaintances, co-workers, and students, and about their beliefs about the characteristics of groups. It does include (in the stereotype sections) research on people’s beliefs about the characteristics of Democrats and Republicans; but it does not address issues of policy or voting.

Prescriptive beliefs (“a woman’s place is in the home”; “children should be seen and not heard”) and moral beliefs (“abortion is immoral”) are often social beliefs in some sense, but they are also beyond the scope of this book. These are beliefs about how people supposedly “should” be, not about how they are. How people “should” be is entirely a matter of opinion, and, therefore, the validity of such “shoulds” is not knowable.

Also, despite the fact that this book does heavily draw on psychological and social scientific research, it does not review every theory or study of expectations or stereotypes. There is so much research on these topics—literally, thousands of studies—that it is not possible to review them all here. Similarly, although this book touches on issues of prejudice and discrimination, it is not fundamentally about prejudice, discrimination, racism, sexism, etc.

Because so many high-quality sources on these issues are cited, I simply urge readers interested in these topics to peruse the reference list at the end of this book.

This book also does not address the self—no self-esteem, self-perceptions, self-beliefs, self-efficacy, self-schemas, and the like. Beliefs about the self are mostly off topic for this book. This book is about people's beliefs about other people, not about themselves. Issues of accuracy, bias, and self-fulfilling prophecy do come up with self-beliefs, but they are sufficiently rich and complex that to address them would require another entire book.

One last “not included.” This book rarely discusses empirical research that did not directly assess accuracy, bias, or self-fulfilling prophecy. Both laypeople and researchers often reach unjustified conclusions on the basis of research that sort of implies things, without actually testing them. This comes up quite specifically with research on interpersonal expectations and is addressed head-on throughout this book. For example, as discussed in Chapters 3, 6, and 10, the early research on interpersonal expectations was often interpreted as suggesting that such expectations were widely inaccurate, even though the early research never assessed the accuracy of those expectations.

Similarly, people have often reached conclusions about the power of error, bias, and self-fulfilling prophecy on the basis of research that has not addressed those issues. Perhaps this can best be illustrated with an example. A few years ago, I gave a talk on research regarding the accuracy of social stereotypes, and concluded that there is more accuracy in stereotypes than social scientists usually acknowledge. At the end of the talk, a young woman who was at Stanford's social psychology program came up to me, all worked up, because my talk did not address “stereotype threat.”

Stereotype threat was originally the idea that fear of confirming a stereotype leads African Americans and women to underachieve on standardized tests (Steele, 1997; it has subsequently been expanded to include all sorts of fears of confirming all sorts of stereotypes). Typically, the research involves making salient (or not salient) either the stereotyped group (African Americans, women) or the supposed area of group vulnerability according to the stereotype (intelligence test performance or math test performance, respectively, for African Americans or women), and then showing that the group performs worse when either the stereotype or the vulnerability is salient.

This was interesting and creative work, but where was the accuracy assessment? That is, whose stereotypes were assessed and shown to be wrong? No one's. Stereotype threat research never addressed the accuracy of any particular person's or group's stereotypes. It assumes only that people fear confirming what they believe to be cultural stereotypes about their groups. Determining the extent to which any individual (or large groups of individuals) actually holds those stereotypes was never the point of stereotype threat research, so it was never assessed. If Sean's beliefs about whether it is going to rain tomorrow are not assessed, we cannot conclude anything about the accuracy of his belief about the weather. If Omar's beliefs about African Americans are not assessed, we cannot conclude anything about their accuracy. Stereotype threat research is interesting and important, but it provided no evidence regarding accuracy. Thus, I did not include it in my talk, and it will not be discussed when I address the accuracy of stereotypes (Chapters 15 through 19).

Stereotype threat is, however, a prime example of research that can be used to suggest or imply something that it does not actually justify (e.g., “stereotypes are inaccurate”). As such, it is a prime example of research that kinda sorta could be seen as relevant to the issues

addressed in this book, but which, because it does not directly address those issues, for the most part, is not included (I touch on it briefly in Chapters 12 and 20). There are, however, many other areas of research besides stereotype threat that, with enough intellectual gymnastics, could be seen as relevant or might actually be relevant in some way, or that seem to suggest something about accuracy, bias, or self-fulfilling prophecy. But, except for research that directly addressed one or more of those phenomena, in this book, they are not likely to be discussed.

Preliminaries

How, then, do social beliefs relate to social realities? Conventional wisdom in much of the social sciences, especially social and cognitive psychology, is that the glass is 90% empty. People are supposedly biased and error-prone, and so many of their beliefs, judgments, and memories are distorted that it is little short of scandalous. I exaggerated for effect there, because no researcher ever wrote anything quite so damning. Whatever exaggeration is there is only slight, though. I am not alone in interpreting psychology as painting this dark picture (e.g., Krueger & Funder, 2004). Furthermore, social and cognitive psychology are replete with research emphasizing error and bias. Psychology's most recent Nobel Prize winner, Daniel Kahneman, received it for demonstrating that people's judgments are subject to an endless slew of biases that lead their conclusions to deviate from expert models of rational decision making and choice. Books emphasizing bias, prejudice, racism, or sexism, or that simply emphasize error (e.g., Gilovich's *How We Know What Isn't So* [1991]; Nisbett & Ross's *Human Inference: Strategies and Shortcomings of Social Judgment* [1980]) vastly outnumber books that address accuracy.

This strikes me as a strikingly odd state of affairs. One common response I often get from colleagues whose own research is on error and bias is that they do not deny the existence of accuracy; they just consider error and bias so much more important because they create so many problems. Sounds good, right? After all, there is tons of scholarship, say, on health problems such as cancer and heart disease. And yet, there is also tons of scholarship on health—hundreds and thousands of books on food, eating well and eating right, exercise, and recreation. Are there really more books on cancer, heart disease, diabetes, and the like than there are on biking, walking, hiking, jogging, tennis, basketball, weight lifting, etc.? Maybe, but there are so many it is hard to know. If so many people truly believe in the importance of accuracy and the strengths of human perception, why is there not more scholarship on these topics? Actually, Chapter 10 provides a whole set of answers to this question, but, at this point, it is really a rhetorical question. My colleagues' protests notwithstanding, I will believe that conventional wisdom in the social sciences accepts the strengths, success, and common accuracy of social perception when the scholarship reflects such wisdom. As of right now, most of it does not.

Table 1–1 presents a short and woefully incomplete list of some of the biases studied and discovered by social and cognitive psychologists. It is an impressive list. And none of them are “false.” They all really exist. Indeed, to this day, one of the shortest routes to success in social and cognitive psychology is to be the discoverer of a new bias. At minimum, though, I think it is fair to say that psychology in particular, but the social sciences more generally,

TABLE 1-1

Social and Cognitive Psychology: Bias after Bias after Bias		
anchoring	base-rate fallacy	biased assimilation
acquiescence bias	social desirability	system justification
conjunction fallacy	ethnocentrism	expectancy bias
false consensus	false uniqueness	availability heuristic
hot hand fallacy	hypothesis-confirming bias	illusion of control
in-group bias	halo effect	illusory correlation
just world bias	linguistic bias	confirmation bias
Fundamental attribution error		
labeling effects	outcome bias	overconfidence
prejudice	pluralistic ignorance	hindsight bias
mindlessness	self-fulfilling prophecy	representativeness
self-serving bias	self-consistency bias	fixed pie bias
unrealistic optimism	out-group homogeneity	belief perseverance
misanthropic bias	stereotype exaggeration	sexism, racism
naïve realism	stereotype-confirming biases	prejudice
social dominance orientation		heterosexism
homophobia	law of small numbers	

have so heavily emphasized error, bias, and irrationality that the message they communicate, at least to a great many people, is that the glass is 90% empty.

How, then, given this overwhelming mountain of data showing that the glass is so very empty, is it remotely possible for anyone, myself included, to come along and suggest, “Nah, the glass is really 90% full”? The answer, as it turns out, takes a whole book.

2 Social Reality Is Not Always What It Appears To Be

THE SCIENTIFIC ROOTS OF RESEARCH ON INTERPERSONAL EXPECTANCIES

MOST PEOPLE WOULD probably agree with the idea that “things are not always what they appear to be.” Unfortunately, it is much more difficult to apply this in one’s daily life. If something appears to us to be some way, we rarely consider the possibility that our perceptions, beliefs, or evaluations may be flawed. After all, we know what we see and hear, right? It is either snowing out or it is not; John is either playing basketball or he is not; and my kids are either being obnoxious or they are not.

It usually is easy to tell whether it is snowing. But the other two examples are less obvious. John may be playing basketball, but his primary interests may be to get exercise or socialize with friends. And what if he is on a team on the basketball floor, but he is just standing around most of the time? Is that playing? And what about my kids? Are they actually being obnoxious, or do I just find them more annoying than usual because I had a long, irritating day at work? Or, worse yet, has my cold, irritable behavior actually turned them into obnoxious little monsters?

The idea that we often assume that our perceptions, judgments, and beliefs are correct, with little or no examination, is called “naïve realism” (Robinson, Keltner, Ward, & Ross, 1995). It is called “naïve” for at least two reasons: (1) It reflects a sort of automatic and innocent faith in the truthfulness of one’s own perceptions, and (2) it reflects a profound ignorance of the ways in which our own actions, motivations, beliefs, expectations, and experiences might shape and influence that which we perceive. The term “realism” refers to the idea that there is an objective reality out there, independent of our subjective perceptions.

Naive realism, therefore, refers to a sort of innocent, unexamined presumption that our perceptions, judgments, etc., are true.

Of course, it may not be so unreasonable for us to believe what we believe—what other choice do we have? Nor is it reasonable to expect people to have much knowledge of 70 years of social psychological research attesting to all the ways in which their heads and hearts influence their perceptions of things in the world. Furthermore, until the 1950s, even most psychological work on perception focused primarily on how people perceive objective stimuli. This “Old Look” in perception emphasized the objective nature of perception (in contrast to the “New Look,” which emphasized subjective influences on perception—discussed later in this chapter). One must remember that, from the 1920s until the mid-1960s, behaviorism overwhelmingly dominated psychological research. And behaviorists (i.e., most of American psychology) banished “internal states” (needs, fears, motives, expectations) from scientific consideration. One could only examine stimuli external to the organism and then assess how the organism reacted to such stimuli.

Nonetheless, one of the most profound contributions of social psychology to understanding the human condition has been the demonstration, time and time again, that our perceptions and judgments, even of events that seem clear and objective to us, may not entirely reflect objective social reality. Instead, they often at least partially reflect our own fears, needs, and beliefs. This was a striking and important discovery of early social psychology, and in this and the next several chapters, I hope to convey some of the excitement and controversy that surrounded those discoveries.

At the same time, however, I also think that it is easy to make too much of the early research and interpret it as demonstrating that internal states (including, but not restricted to, expectancies) produce powerful biases in perception. Few, if any, of the early studies justified strong conclusions. It is one thing to claim that biases creep into social perception (a conclusion that I think is justified); it is quite another to claim or imply that social perception is dominated by bias (a conclusion that I think is not justified).

In this chapter, I review and critically evaluate the early social psychological research demonstrating how people’s motivations and beliefs sometimes bias their perception of seemingly objective physical and social phenomena. There were three separate lines of research that ultimately revolutionized the ways in which psychologists understand social perception: early research on stereotypes, early work on person perception, and Merton’s (1948) classic analysis of self-fulfilling prophecies. Although each area suffered from major flaws or limitations, when taken together, they provided a compelling case that, at least sometimes, the internal states of the perceiver, rather than or in addition to the objective external characteristics of those being perceived, influence just what was being perceived.

By “early research” I generally refer to research on bias and subjective influences on perception that appeared before 1960 (in the 1960s research directly focusing on interpersonal expectancies took off—but that story has to wait till the next chapter). Of course, a comprehensive review of all such research is beyond the scope of this chapter. Therefore, I focus primarily on the broad theoretical perspectives, a small number of studies widely recognized as classics in the area of bias, and a few other studies that aptly illustrate some of the main theoretical ideas that provided much of the foundation for modern approaches to interpersonal expectancies.

Earliest Work on Stereotypes

Lippmann. One of the first arguments that our perceptions are not necessarily strongly linked to objective reality did not even come from a social scientist—it came from a journalist. In a broad-ranging book called *Public Opinion*, Walter Lippmann (1922) touched on *stereotypes*—and defined them in such a way as to color generations of social scientists' views of stereotypes.

Lippmann suggested that people could not understand, remember, or interpret the vast array of stimuli to which they were exposed. To understand the world in its full complexity, he argued, is an impossible task. So, according to Lippmann, they must simplify and reduce the overwhelming amount of information they receive. Stereotypes, for Lippmann, arose out of this need for simplicity. He believed that people's beliefs about groups were essentially "pictures in the head."

The term "picture in the head" may seem reasonable and may capture some truth. It probably is easy for most of us to conjure up clear images of Islamic fundamentalists, British bank executives, Asian engineering students, New York Jews, ghetto Blacks, and French artists. But think about just what a "picture" is. It is a static, two-dimensional representation of a four-dimensional stimulus (most real-world stimuli have width, length, and depth, and also change over time). A picture is rigid, fixed, and unchanging. It is oversimplified, in the sense that a picture can never capture the full complexity of life for even one member of any group.

Lippmann's metaphor (pictures in the head), therefore, created an image of stereotypes as oversimplified, superficial, rigid, fixed, and at least partially out of touch with reality. This should sound pretty familiar—it constitutes the working definition of stereotypes that many people, including many social scientists, still hold today. Thus, it constitutes one of the earliest perspectives suggesting that people's social beliefs may not be fully in touch with social reality.

Katz and Braly. This was perhaps the first empirical study of stereotypes. Katz and Braly (1933) were interested in discovering what people thought were the main characteristics of various national, racial, and ethnic groups. They had 100 Princeton students assign traits to 10 ethnic groups. Their results were striking—there was widespread agreement on the most prevalent traits for each group. For example, about 80% of their students described Germans as scientific, Blacks as superstitious, and Jews as shrewd. Fifty-four percent described Turks as cruel, even though not a single Princeton student had ever met a Turk!

These levels of agreement were so high that Katz and Braly could not believe they were even remotely accurate—more likely (Katz and Braly believed), they reflected the preexisting biases, stereotypes (in the Lippmann sense), and expectations of the students. Their argument had essentially two components. First, even if there are differences between groups, there will also be a great deal of similarity or overlap between them. Even if Germans as a nationality were more scientific minded than, for example, Italians, there are tons of nonscientific Germans and many scientific-minded Italians.

Second, they seemed to assume that nonprejudiced responses would be based on an objective assessment of personal experiences (this is implied in their writing—they never say this in so many words). If so, then because of their first argument (lots of overlap among groups),

almost as many people should have experience with nonscientific Germans and scientific Italians as with scientific Germans and nonscientific Italians. Thus, although there might be some difference in the percentage of people identifying Germans and Italians as scientific, those differences should not remotely approximate the huge differences they actually found.

This analysis has several severe logical and empirical flaws (Jussim, McCauley, & Lee, 1995; McCauley, Jussim, & Lee, 1995; McCauley, Stitt, & Segal, 1980; see also Chapters 11, 15, and 16). For now, I will simply point out that (1) their assumption that only beliefs based on direct personal experience can be accurate is most odd (although it pervaded the stereotyping literature for years!), and (2) the mere fact of high agreement does not preclude high accuracy. Indeed, there are many reasons for considering high agreement as one (of several important) criterion for inferring high accuracy (e.g., Funder, 1987, 1995; Kenny, 1994; see also Chapter 11). The idea here is simple: If everyone is accurate, they must all agree. Thus, in general, agreement is a good, if imperfect, indicator of accuracy (they could all agree and still be inaccurate).

Katz and Braly's (1933) discussion of stereotyping particularly contributed to the burgeoning consensus that stereotypes biased social perception and perpetuated social injustice. In their discussion section (p. 288), they provided a description of the role of stereotypes in person perception that could have appeared in the discussion section of many modern articles on stereotyping:

Of course, individual experience may enter into the student's judgment but it probably does so to confirm the original stereotype which he has learned. . . . When he meets a German, he will expect the scientific trait to appear, and because human beings from time to time exhibit all kinds of behavior he can find confirmation of his views. . . . [When] people with a prejudice against Jews . . . meet a flagrant contradiction of their stereotyped picture in a specific Jewish acquaintance . . . they observe that this Jew is an exception. . . . By thus omitting cases which contradict the stereotype, the individual becomes convinced . . . that its members are just the kind of people he always thought they were.

Although Katz and Braly (1933) actually presented no evidence to support this analysis, their emphasis on inaccuracy and irrationality in stereotypes was highly influential and helped set the tone for the next 60 years of research on stereotypes (see, e.g., Allport, 1954; Ashmore & Del Boca, 1981; Brigham, 1971; Lee, Jussim, & McCauley, 1995). This work also set the stage for the next early classic of stereotyping research—one which seemed to provide much clearer evidence of inaccuracy and irrationality in stereotypes.

LaPiere (1936). An early attempt to more clearly document inaccurate and irrational stereotypes was a study of beliefs regarding Armenians living in California. Of all the oppressed and stigmatized groups in America, why Armenians? I do not know. Apparently, however, in the 1930s in California, there was quite a lot of hostility toward Armenians.

LaPiere (1936) first interviewed over 600 non-Armenians and found that *none* approved of family members marrying Armenians, most would not want to belong to clubs including Armenians, and only a minority said they would even work with them. Thus, it is clear that many locals did not like Armenians. However, that pretty much ends LaPiere's summary of his interviews.

Note the complete lack of information on people's *beliefs* about Armenians. This is important because dislike and inaccuracy are not the same thing. For example, I deeply dislike Ku Klux Klan members, in large part, because I think they despise minorities and because they sometimes express their beliefs through violence. Are my beliefs about Klan members (i.e., my stereotype) wrong? I do not think so. So if one can despise a group and still perceive them accurately, researchers cannot justifiably infer inaccuracy of beliefs about a group when they only demonstrate dislike of that group.

The Klan is easy, but what about ethnic groups? Can't we assume that dislike of an ethnic group reflects biased and irrational thinking? Not necessarily—this fundamentally depends on whether the dislike is justified. Can disliking an ethnic group ever be justified? If Bosnians despised Serbians during the Yugoslav civil war because they believed the Serbians engaged in mass murder of Bosnians, would the Bosnians have been wrong? I do not think so. It is not hard to generate numerous examples like this. Dislike and inaccuracy do not always or necessarily go hand in hand. Thus, LaPiere's (1936) demonstration of widespread dislike of Armenians did not demonstrate inaccuracy. In fact, LaPiere (1936) presented only sparse information regarding respondents' *beliefs* about Armenians—and only the beliefs, not liking or disliking, can be examined for their accuracy.

Throughout the rest of his article, LaPiere (1936) did present anecdotal evidence regarding inaccuracy. For example, he claimed (p. 233) that "The most frequently advanced explanation for antipathy towards the Armenians of Fresno County is that they are 'dishonest, lying, deceitful.'" By whom? His respondents? He did not say so. And how frequently is "most frequently"? We do not know because he did not tell us. He did go on to quote single individuals throughout the rest of his paper. For example, he quoted (p. 233) a local credit officer as claiming that when it comes to credit, "the Armenians are, as a race, the worst we have to deal with." Although this is a very striking quote, the extent to which this one credit officer's statement reflected the views of the other 599 people in LaPiere's survey is unknown, because it was not reported.¹

He also claimed that 30% of his sample believed Armenians were "parasitic"—that is, they were not self-sufficient, so became a disproportionate financial burden on the community. Of course, that means that 70% of his sample *did not* make this claim. Nonetheless, LaPiere went on to quote (p. 234) a local hospital official as claiming that Armenians were "constantly demanding more charity here than do other races."

LaPiere also claimed that there was a widespread belief that Armenians were troublemakers who frequently got into trouble with the law. He claimed that this belief was particularly widespread among public officials and nearly all the lawyers he interviewed, but, again, he provided no numbers or details.

In short, then, LaPiere claimed that there were three main charges the locals leveled against Armenians: They engaged in shady business dealings resulting in poor credit, they excessively relied on public charity, and they frequently were in trouble with the law. In the rest of his article, LaPiere went on to document the invalidity of these beliefs. With respect to understanding just how inaccurate his respondents were, it is therefore unfortunate that his reporting of their actual beliefs was so sketchy. Furthermore, at least one subsequent researcher (Mackie, 1973) obtained a copy of LaPiere's dissertation (on which the published study was based) and concluded that there was considerably more evidence of accuracy than his article suggested.

But let's leave all that aside. For now, I am simply documenting the history of research that led social scientists to conclude that people's perceptions are often biased and inaccurate. So let's stipulate for the moment that LaPiere was right on all three counts—lots of non-Armenians believed that Armenians were sleazy, dependent on the public dole, and immoral.

One unique contribution of LaPiere's (1936) study was the use of objective criteria to examine the validity of each of the accusations leveled against Armenians. He obtained public records to examine the truth of each charge. First, he found that local banks considered Armenians nearly as creditworthy as non-Armenians. Nearly equal numbers of Armenians and non-Armenians were considered bad risks; more Armenians were considered fair risks; more non-Armenians were considered excellent risks. This does seem to provide at least some evidence that Armenians were not as good credit risks as non-Armenians. But the difference was not huge. Furthermore, the evaluations of creditworthiness were provided by non-Armenian bank officials, among whom there was at least some potential for bias. Thus, this pattern did not seem to justify the belief that Armenians were routinely shady businesspeople with poor credit histories.

His evidence for the inaccuracy of beliefs regarding the next two accusations—public burden and immoral—was even clearer. He examined both hospital records and local welfare records and found that Armenians requested charity or public assistance at about 15% to 20% the rate of the non-Armenian local population. They were dramatically *less* likely to become a public burden than were non-Armenians. Similarly, LaPiere (1936) examined court cases to assess involvement of Armenians in legal troubles. He consistently found that Armenians were about one-quarter to one-third as likely as non-Armenians to appear in court cases. Again, Armenians seemed dramatically *less* likely to have legal difficulties than did their non-Armenian neighbors.

LaPiere concluded that people's explanations for why they disliked Armenians were entirely bogus—they were rationalizations concocted to justify their own prejudice. That is, the prejudice comes first, and *causes* people to develop specious and false negative images of the despised group. Thus, people saw sleaziness, dependency, and immorality, not because Armenians actually had these attributes, but to justify their own antipathy toward Armenians.

G. Allport (1954/1979). Gordon Allport's classic book, *The Nature of Prejudice*, provided a broad and sweeping analysis of stereotypes, prejudice, and discrimination. This book was so influential that it set much of the research agenda on stereotypes and prejudice for the next 50 years and remains widely cited today. Allport reviewed and systematized many of the themes that first appeared in Katz and Braly (1933) and LaPiere (1936) and, indeed, in much social science work on prejudice up to that time. G. Allport (1954) distinguished between, on the one hand, rational and flexible beliefs about groups, and, on the other, stereotypes. For G. Allport, stereotypes are faulty exaggerations. All-or-none beliefs, such as "all Turks are cruel" or "all professors are absentminded," are stereotypes that are clearly inaccurate, overgeneralized, and irrational, because there are virtually no social groups whose individual members universally share some set of attributes. G. Allport also characterized stereotypes as unjustifiably resistant to change, steeped in prejudice, and concluded they were a major contributor to social injustice.

G. Allport also argued that stereotypes led to all sorts of biases and errors in social perception. In one of his own studies, G. Allport (1954/1979) showed people a picture of an African American man in a business suit and a White man holding a razor. Later, when asked to

describe the picture, many “remembered” the African American holding the razor and the White wearing the business suit! This is a stereotype-based bias in person perception similar in tone and upshot to many of the biases LaPiere (1936) found regarding Armenians as a group. Allport also reviewed numerous studies and presented many amusing anecdotes demonstrating the ways in which stereotypes and prejudice undermine the objectivity of social perception. Thus, it was clear even by the 1950s that, at least sometimes, people’s beliefs about groups, and their perceptions of individuals from those groups, were not always based in objective social reality.

Early Social Perception Work

The “New Look” in perception. “OK,” I can almost hear you say, “I concede that the beliefs held by bigots are not necessarily particularly objective. But that does not mean the rest of us go through our days allowing our preconceived notions, needs, and desires to unduly color our interpretations of the world.” What about the rest of us? How much are our perceptions biased by our own expectations and motives? This was precisely the question addressed by the revolutionary “New Look” in perception of the 1940s and 1950s.

The New Look was, in large part, a response to and reaction against the prevailing “Old Look” in perception. Prior to the New Look, work in perception emphasized the objective aspects of perception. Work focused on the neural reception of external stimuli, on psychophysics (e.g., mathematical models of stimulus detection), and on the registration of stimuli by the brain. The dominant behaviorist perspective of the period banished phenomena such as fears, needs, and expectations from study, dismissing such internal states as unscientific.

Then came the New Look researchers who, en masse, said, “Whooooaaa! You can’t banish needs, motives, and expectations—they all can play a *crucial* role in determining just what is perceived in the first place!” And demonstrating this became a major goal of research for social and personality psychologists in the 1940s and 1950s. The main claims of the New Look could be captured by two terms: perceptual vigilance and perceptual defense. Perceptual vigilance referred to the tendency for people to be hypersensitive to perceiving stimuli that met their needs or were consistent with their values, beliefs, or personalities. Hungry people, for example, should be more likely to detect food and perceive ambiguous stimuli as food.

Perceptual defense referred to the tendency for people to avoid perceiving stimuli that was uncomfortable or threatening. Imagine yourself sitting around a Thanksgiving dinner table with your extended family, and your favorite aunt starts discussing having sex with your uncle. At first, you would probably not believe your ears—you might even misinterpret her as saying she was “in a real good funk with your uncle.” If you had such a reaction, you would have experienced something akin to perceptual defense—the reluctance to perceive stimuli that might make you feel uncomfortable.

F. Allport (1955) summarized the New Look research as emphasizing and *seeming* to support six hypotheses: (1) Bodily needs determine what is perceived; (2) reward and punishment determine what is perceived; (3) personal values facilitate recognition of words related to those values; (4) the value of objects to an individual influences their perceived magnitude; (5) people perceive things in a manner consistent with their own personality

characteristics; and (6) people take longer to perceive disturbing stimuli (than pleasant or neutral stimuli), and such disturbing stimuli evoke emotional reactions even before they are perceived.

Taken together, the body of research examining the New Look seemed to provide impressive and convincing evidence regarding the role of needs, motives, etc., in the perceptual process. But not so fast—F. Allport (1955) also provided a thoughtful critical evaluation of the research supporting each of the six hypotheses. In general, he concluded that the actual evidence for the New Look proposals was inconclusive—they *could* be right, but the particular studies often did not provide clear and convincing evidence. I will briefly discuss two examples to convey a sense of F. Allport's analysis.

A study by Levine, Chein, and Murphy (1942) examined the hypothesis that bodily needs determine what is perceived. Specifically, "we determined to measure the relation between the intensity of the food interest and the amount of perceptual distortion" (p. 286). They did this by (1) depriving participants of food for 1, 3, 6, or 9 hours; (2) showing them pictures behind a ground glass screen, which rendered them difficult to see clearly; and (3) asking them to "verbalize an association with every picture you see" (p. 289). They examined these associations to see how often participants mentioned food.

The results seemed to strikingly confirm the hypothesis that hunger influenced perception. The frequency with which participants associated food with the pictures increased through the first 6 hours of food deprivation (although there were fewer food associations after 9 hours than after 6 or even 3 hours). Hungrier people seemed to actually see more food!

However, as F. Allport pointed out, there are at least two major flaws with this study. First, the decrease in food associations after 9 hours of deprivation is not consistent with the hypothesis. Second, and even more important, is that associating food with a drawing is not the same as perceiving the drawing as depicting food. People were not asked to describe or identify the drawings. They were only asked to come up with associations with the drawings. If you show me a picture of a baseball game and ask me to come up with associations, and I say "hot dogs," that does not necessarily mean that I perceive hot dogs. The finding that food associations increased through up to 6 hours of food deprivation is interesting. Nonetheless, their explicit and repeated claim that they were concerned with assessing the effects of bodily needs specifically on perception notwithstanding, it is not clear that this study provides any information about perception.

Perhaps the most controversial results concerned perceptual defense. Several studies all used essentially the same approach, testing for perceptual defense in similar ways (see F. Allport, 1955, for the details). The main hypothesis was that it takes longer to perceive threatening words than nonthreatening words. Researchers typically presented words tachistoscopically (a tachistoscope is a machine that can present stimuli, such as slides, at varying rates of speed). They would start by presenting words so quickly that no one could recognize them (e.g., at 1/100th of a second). Then they would slowly increase the exposure time until all the words were recognized. The key experimental manipulation was the type of word. Some were neutral, everyday words, but others were unpleasant or taboo. In many of the studies, they also assessed subjects' galvanic skin response (GSR; this measures the electrical conductivity of the skin—which reflects anxiety or tension).

A study by McGinnies (1949) was typical. With a tachistoscope, people were presented with neutral words (such as "apple," "child," and "glass") or taboo words (such as "bitch,"

“whore,” and “raped”). Their galvanic skin response was also measured. The results consistently were quite striking: (1) It took longer for people to recognize taboo than neutral words; (2) prior to recognition, galvanic skin responses were higher for threatening than for neutral words (suggesting unconscious anxiety); and (3) neutral words were most likely to be perceived as other neutral words, but threatening words were more likely to be perceived as neutral words or as nonsense words. This seemed to provide compelling evidence for perceptual defense.

The evidence, however, was less compelling than it seemed at first glance. There were two main empirical difficulties with studies such as McGinnies (1949). First, it may have taken longer for subjects to recognize taboo words, not because they were emotionally threatening, but simply because they were unusual. It is often easier to see, hear, or recognize things we expect to find, and it often may just take a while to figure out the meaning of something completely unexpected. Furthermore, words such as “whore” and “penis” were relatively uncommon, and certainly unexpected, in a research lab in the 1940s. Thus, it may have taken longer to recognize the taboo words, not because they were threatening and defended against, but only because they were unexpected.

Second, however, although this alternative does argue against the perceptual defense explanation offered by the New Lookers, it still argues for an important role of an internal state in perception. Specifically, it becomes easier to find what one expects than what one does not expect. Whether or not this explained the New Look findings, the idea that, at least sometimes, expectations influence and bias perception has generated a vast amount of research and eventually became widely accepted.

Returning to the New Look studies, F. Allport also highlighted a second difficulty that was even more damaging to the perceptual defense claim. Perhaps the longer time to report recognition of words like “penis” and “whore” had nothing to do with perception—instead, perhaps delayed reporting occurred because subjects were embarrassed and reluctant to report seeing such words in the context of a staid and professional research laboratory experiment. This would also account for the occurrence of higher GSR scores (indicating higher anxiety) prior to recognition of the taboo words. F. Allport doubted that methodological procedures existed that could adequately address this problem.

Hastorf and Cantril (1954). This study is not usually discussed with the New Look, but I discuss it here because it is such a great example of the role of prior beliefs and motives in social perception. In 1951, Dartmouth and Princeton played a hotly contested, aggressive football game. A Princeton player received a broken nose; a Dartmouth player broke his leg. Accusations flew in both directions: Dartmouth loyalists accused Princeton of playing a dirty game; Princeton loyalists accused Dartmouth of playing a dirty game.

Into this mix stepped Hastorf and Cantril (1954). First, they surveyed students of both schools and, not surprisingly, found that Princeton and Dartmouth students had different opinions about the game. But they also provided a more direct test of whether the students were actually “seeing” different games. They showed a film of the game to 48 Dartmouth students and 49 Princeton students and had them rate the total number of infractions by each team. Dartmouth students saw both the Dartmouth and Princeton teams as committing slightly over four (on average) infractions. The Princeton students also saw the Princeton team as committing slightly over 4 infractions, but they also saw the Dartmouth team as committing nearly 10 infractions. This study has long been cited as a demonstration of how

motivations and beliefs color social perception (e.g., Schneider, Hastorf, & Ellsworth, 1979; Sedikides & Skowronski, 1991). Hastorf and Cantril (1954) themselves concluded that Princeton and Dartmouth students seemed to be actually seeing different games.

The New Look in retrospect. Despite the flaws and limitations of their actual studies, and despite losing many intellectual battles with those challenging their interpretations at the time, the New Lookers ultimately won the war—and the victory was nearly absolute. Within social and personality psychology, the idea that motivations, goals, and expectations influence perception is now so well-established that it is largely taken for granted. Since 1970, research on social perception has been dominated by an emphasis on the many errors and biases that characterize social perception (e.g., Fiske & Neuberg, 1990; Fiske & Taylor, 1991; Gilovich, 1991; Greenwald & Krieger, 2006; Jones, 1990; Kahneman & Tversky, 1973; Myers, 1999; Nisbett & Ross, 1980—some of this work will also be discussed in detail later in this book). In fact, the shift away from an assumption of accurate perception of stimuli and toward an emphasis on inaccuracy and bias was so nearly absolute that virtually no accuracy research was conducted in social psychology for about 30 years, and some of the most prominent researchers of the day pronounced it to be essentially an uninteresting, dead topic, and one which was very difficult to study anyway (e.g., Jones, 1985, 1990; Schneider et al., 1979).

F. Allport (1955) saw this coming. The New Look researchers strongly and consistently emphasized subjective influences on perception and de-emphasized (or just did not study) objective influences on perception. F. Allport was concerned, therefore, that social and personality psychologists were, intentionally or not, conveying a vision of social perceivers so dominated by their own needs, motives, values, and beliefs that they were out of touch with reality. Furthermore, he was concerned that this would create an overly pessimistic and largely inaccurate theoretical perspective on the nature of social perception:

Where the perception is bound so little by the stimulus and is thought to be so pervasively controlled by socially oriented motives, roles, and social norms, the latitude given for individual and group differences, for deviating and hence non-veridical awareness, is very great. (p. 367)

He also warned against overemphasizing bias and inaccuracy:

What we are urging here is that social psychologists, in building their theories of perception, assume their share of the responsibility for reconciling and integrating their “social-perceptual” concepts, fraught with all their deviations and special cognitive loadings, with the common and mainly veridical character of the basic human perceptions. (p. 372)

F. Allport was right on both counts—his concern that the New Look could lead to an overemphasis on subjective influences on perception could not have come more true; and he was right to urge social psychologists to develop theories that presented a more balanced vision of the roles of error, bias, and accuracy in social perception.

As a case study of both of F. Allport’s points, consider the Hastorf and Cantril (1954) football study described previously. This is one of the few studies from this era that has endured—it has been regularly cited as an example of the nonveridical, biased, and socially

constructed nature of social perception (e.g., Schneider et al., 1979; Ross, Lepper, & Ward, 2010; Sedikides & Skowronski, 1991). As far as I know, it has never been cited as a testament to the “largely accurate nature of the basic social perceptions.” In fact, however, I think this, far more than bias and inaccuracy, is the basic message of the *results* of the study (even though it was not the authors’ message).

Because they provided no objective assessment of infractions, we must use agreement as a reasonable proxy for accuracy (e.g., Funder, 1987, 1995; Kenny, 1994; see Chapter 11). This fits well in most sporting events. If my team and your team both agree that “the call” (the ball was in, it was out, it was caught, it was missed, you were offside, I illegally held, etc.) was correct, there is no controversy and, for all practical purposes, the call is correct.

To get a handle on how much agreement there was, let’s consider the level of agreement in each cell of their design. There were four cells (Princeton students’ perceptions of the Princeton team, Princeton students’ perceptions of the Dartmouth team, Dartmouth students’ perceptions of the Princeton team, and Dartmouth students’ perceptions of the Dartmouth team).

Perfect agreement would mean that Princeton and Dartmouth students always agreed with each other. Let’s just focus on the type of information Hastorf and Cantril focused on: perception of numbers of infractions. For example, let’s say both groups believed that Dartmouth committed five infractions. This would represent perfect agreement regarding the number of Dartmouth infractions. Using their standards (which I would not actually use for reasons described later), if one group saw 10 infractions and the other 0, that would be complete disagreement.

Were students’ perceptions dominated by bias? Well, it depends on what the phrase “dominated by bias” means. To me, “dominated by bias” certainly means “more bias than accuracy,” and that is how I will use it in the remainder of this analysis.² This leads to a very simple and straightforward criterion for evaluating whether the Princeton and Dartmouth students’ perceptions were dominated by bias or by accuracy (keeping in mind that the only criterion for accuracy in this study was agreement): (1) The “dominated by bias” hypothesis predicts that there should be less than 50% agreement in perceptions of infractions; (2) the “dominated by accuracy” hypothesis predicts that there should be more than 50% agreement in perceptions of infractions.

Now let’s evaluate Hastorf and Cantril’s (1954) actual results. First, the judgments regarding the Princeton team were nearly identical (both groups saw about four infractions). Thus, for the half of the judgments regarding the Princeton team, there was nearly 100% agreement.

Now let’s consider judgments regarding the Dartmouth team. The Princeton students saw the Dartmouth team commit 10 infractions; Dartmouth students saw Dartmouth commit 4 infractions. This is 40% agreement.

There are several ways this can be evaluated. First, we can simply average the percentage agreement. For half the judgments, there was 100% agreement; for the half there was 40% agreement. Averaging over the two sets of judgments, therefore, shows that overall agreement was 70%.

Even this, however, substantially underestimates level of agreement, because Hastorf and Cantril (1954) did not report the total number of plays that appeared on the film that they showed. Unfortunately, I was unable to track down a box score for this game. Nonetheless, a

typical college football game has 60 to 100 plays. Indeed, the 60 figure is extremely low. But let's work with this lower figure because use of the higher figure would lead to a conclusion of even more agreement.

One aspect of my analysis so far might appear to overestimate agreement. Just because Princeton and Dartmouth students agree on a certain overall number of infractions does not necessarily mean that they saw the *same* infractions. Therefore, let's work with the worst-case scenario for accuracy—they disagreed on *all* perceptions of infractions. This would mean that there were, at most, 22 disagreements (Dartmouth students saw 4 Dartmouth infractions and 4 Princeton infractions; Princeton students saw 10 Dartmouth infractions and 4 Princeton infractions; all are disagreements; so $4 + 4 + 10 + 4 = 22$ disagreements).

Of course, this all means that there was 100% agreement on the absence of infractions for every play above those 22. In other words, if there were only 60 total plays (and there were probably more), there was 100% agreement on at least 38 plays. That means there was at least 63% agreement.

Of course, even if disagreement was 37%, *bias* was much lower. Why? Because bias in favor of one's college explained, on average, only 3 of 60 judgments for each group of students. Where does this 3/60 figure come from? The groups did not differ in their perceptions of infractions by the Princeton team. Therefore, even if they disagreed regarding when Princeton committed the four infractions, there was no *bias*.

They did differ by six in perceived infractions regarding Dartmouth. This means that there was an average bias of three perceived infractions for students from each college (six infractions divided by two groups of students = three per group). Bias totaling 3 infractions out of 60 plays? This does not seem to confirm a domination by bias hypothesis.

In fact, however, agreement in social perception is probably grossly underestimated by all of my above estimates. Why? Because there are 11 players on each team playing at a time. An infraction occurs when any *one* player commits an illegal act. So, if I say "Princeton's linebacker committed a late hit" (this is illegal), even if you disagree, *we are, in fact, agreeing 10 of 11 times, by virtue of neither of us seeing any of the other 10 Princeton players commit an infraction*. Indeed, we would be agreeing 21 out of 22 times, if neither of us believed Dartmouth committed an infraction on that play.

Hastorf and Cantril (1954) is a great study, partially because it is an early demonstration of bias, partially because they examined bias in a very rich, real-life context, and partially because this context is one which most of us who have ever attended college or been any type of sports fan can readily relate to. So this study, like much of the New Look research, did provide some evidence of bias and subjectivity. But, just as F. Allport (1955) would have suggested, over the decades it has been *interpreted* as emphasizing the biased and idiosyncratic nature of social perception. But if one takes a deeper look at their data, and the context in which their data were gathered, it is clear that they obtained far more evidence of agreement than of bias, even in a hot, emotionally charged context. As F. Allport might have said, this is a testament to "the common and mainly veridical character of the basic human perceptions" much more than to bias.

Asch and implicit personality theory (1946). In another classic piece from this period, Asch (1946) dispensed with the accuracy issue altogether. He performed 10 experiments examining *processes* of social perception, with little apparent concern for issues of accuracy and bias. Indeed, it may not be apparent at first how this research is relevant to interpersonal expectancies (please

bear with me). Asch's (1946) main objective was to demonstrate that people arrive at an overall impression of a person by integrating specific behaviors, traits, and attributes in unique ways that are not readily predictable from their judgments of each specific behavior, trait, or attribute.

His studies are best remembered (and most frequently cited) for their demonstration of the existence of central versus peripheral traits. Central traits exert a pervasive and profound influence on an overall impression of a person; peripheral traits exert relatively minor influences. Warm/cold was a central trait Asch (1946) discovered. When people arrived at an impression of a target described (by the experimenter) as intelligent, skillful, industrious, determined, practical, cautious, and *warm*, they arrived at a very different impression than when they believed the target was identical, except that they were *cold*. The warm targets were seen as much more generous, wise, happy, good-natured, humorous, sociable, and popular. When the *warm/cold* manipulation was replaced by *polite/blunt*, there was little difference in impressions. Thus, *warm/cold* was a central trait because it exerted a pervasive influence on the impression of a target person; *polite/blunt* was peripheral because it exerted a far more modest influence.

This pattern means that people seem to believe that certain clusters of traits hang together. If you know that someone is warm, you infer generosity, humor, etc. These beliefs are often *implicit*—people's judgments and evaluations of others reflect their own beliefs that certain traits hang together, even if they are not consciously aware of this. This phenomenon became known as *implicit personality theory*. It is as if people have their own implicit "theories" regarding the meaning of different personality traits.

Now, we are finally getting closer to expectancies, social perception, and bias. Implicit personality theories can be considered a type of expectancy phenomenon. If "warmth" implies wisdom, then people will expect a person known to be warm also to be wise.

Kelley (1950). This type of analysis quickly led to one of the early classics demonstrating that expectancies can indeed bias person perception. Kelley (1950) introduced a guest lecturer to each of several classes of undergraduates. All students were informed that people who knew the lecturer considered him to be "industrious, critical, practical, and determined." Half were informed that they also considered him "cold;" half were informed that they also considered him "warm." The guest lecturer then led a 20-minute discussion.

This, therefore, is perhaps the first study directly and explicitly examining the role of expectancies in biasing person perception. So what happened? Did the warm versus cold expectation influence the students' judgments of the guest speaker? It sure did. Compared to students who expected the speaker to be cold, students who expected him to be warm judged him as more considerate, informal, sociable, popular, good-natured, humorous, and humane. Of course, we know that he did not actually act differently to the different students. Why? All students saw the exact same guy engage in the exact same behaviors at the exact same time!

Unfortunately, however, there is an important limitation to this study that qualifies the extent to which it demonstrates the power of expectancies to bias students' judgments. The effects on judgment may not have occurred because the warm/cold manipulation led students to actually see (interpret, perceive, judge) the actions and behaviors of the guest lecturer differently. The effect of the warm/cold variable may have occurred directly on the judgment, without affecting students' perceptions of the lecturer behavior. That is, perhaps

both groups of students viewed his lecture identically—but one group had reason to believe he was cold and the other had reason to believe he was warm (i.e., the experimenter told them so). Knowing that he is cold might lead to a lower rating on, for example, sociable and good-natured, regardless of how he acted in the lecture.

The purpose of Kelley's (1950) study was to assess some of Asch's ideas in a real-life context. Its purpose was not to assess accuracy or agreement in students' perceptions of the lecturer. Of course, without an accuracy assessment, the study cannot possibly provide any direct information regarding the *relative* roles of accuracy and bias in social perception. Thus, although it is clear that the warm/cold variable influenced judgments, it is impossible to compare the extent of bias to the extent of accuracy in this study.

The Self-Fulfilling Prophecy

Thus far, I have reviewed the early research on stereotypes and on the New Look that at least raised the possibility that internal states influenced how people perceived external social stimuli. The self-fulfilling prophecy, however, is in some ways a much more radical and extreme notion—the main idea is that people's beliefs can have a profound influence not only on what they *perceive* but also on the actual *behavior* of the people they are perceiving! The self-fulfilling prophecy refers to a false belief that leads to its own fulfillment. Self-fulfilling prophecies influence social perception, not by altering or biasing the perceptual processes of perceivers—but by influencing that which is perceived.

The Last National Bank. Merton (1948) developed the idea of the self-fulfilling prophecy and illustrated it with several examples. One was a parable depicting a common event during the Great Depression. Banks make money by placing people's savings deposits in income-producing investments—mortgages, commercial loans, etc. Thus, although a successful bank will keep a substantial amount of cash on hand to pay depositors, to make a profit, it must invest most of depositors' money.

The Last National Bank was a profitable institution. It remained thriving even into the early years of the Great Depression. But then banks began to fail. Somehow, a rumor got started that the Last National Bank was on the verge of insolvency. Depositors flocked to the bank to salvage their savings before the bank went under. Because the Last National Bank did not (and could not!) keep most of its depositors' savings in cash, it could not pay them all. That is, it became insolvent. The originally false definition of the situation—that the bank was insolvent—had become true. This is the essence of the self-fulfilling prophecy.

African Americans and labor unions. Merton (1948) raised the issue of the self-fulfilling prophecy, however, because he believed it played a profound role in social injustices and inequality. Merton's (1948) first example involved African Americans and labor unions. In the early part of the 20th century, most labor unions barred African Americans from membership. Union members often claimed that African Americans were strikebreakers and could not be trusted. Keep in mind that we are talking the early 20th century—America had recently become an industrial powerhouse, and labor unions were far more powerful than they are today.

Exclusion from the unions, therefore, severely limited African Americans' job opportunities. When faced with a strike, however, companies often offered jobs to all takers, and

African Americans often jumped at the chance for work. Thus, White union members' beliefs about African Americans were confirmed—they really became strikebreakers. But how do we know this is a self-fulfilling prophecy? After all, maybe the White union members' views were correct.

We know they were incorrect from the historical record. At the time Merton (1948) wrote his piece, some unions had begun to admit African Americans, although many still excluded them. Among those that admitted African Americans, strikebreaking was no more common than among Whites. And today, this stereotype of African Americans as strikebreakers is probably so alien to most readers of this book as to seem patently antiquated, quaint, and silly. Why? Because African Americans have not been excluded from unions for years, and there is no longer a shred of evidence that they are disproportionately likely to be strikebreakers. Thus, the widespread belief in the early part of the 20th century was clearly initially false, although at the time, it became true—a classic self-fulfilling prophecy.

Damned if they do and damned if they don't. Merton (1948) also claimed that out-groups were “damned if they did and damned if they did not.” Again, he started with a discussion of African Americans that is appallingly relevant today, over 60 years later (Merton, 1948, p. 200): “Thus, if the dominant group believes that Negroes are inferior, and sees to it that funds for education are not ‘wasted on these incompetents’ and then proclaims as final evidence of this inferiority that Negroes have proportionately ‘only’ one-fifth as many college graduates as whites, one can scarcely be amazed by this transparent bit of social legerdemain.” Thus, the self-fulfilling belief in African American inferiority contributed to a system that undermined African Americans' educational opportunities. But people at the time could point to the “objective” evidence of African American inferiority as justification for continuing such policies. This, then, is an example of an out-group that was “damned if they did not” (possess in-group virtues).

In contrast to the union example, the twisted thing about this one is that it is still largely true today. Most states fund public schools through local property taxes. Of course, this means that wealthier communities usually provide more money for schools than do poorer communities. Because of continued residential segregation, and because, on average, African Americans still earn so much less than do Whites, this policy obviously is discriminatory. In New Jersey, for example, repeatedly over the last 30 years, the State Court has ruled that funding public schools with local property taxes is illegal because it is discriminatory, and each time, the state lawmakers developed some sort of minimalist reform, which did little to reduce the inequities, and which returned the issue to the courts (*New York Times*, 1994).

How about “damned if they do”? Was Merton (1948) really claiming that in-groups derogated out-groups for possessing in-group *virtues*? Yes, he was. He used the example of Abe Lincoln, Abe Cohen, and Abe Kurosawa. If Lincoln worked through the night, it testified to his industriousness and perseverance. When Jews or Japanese did so, they were accused of a sweatshop mentality and of engaging in unfair business practices. Whereas Abe Lincoln was seen as frugal and thrifty, Abe Cohen was seen as stingy. If Abe Lincoln was honored for having been smart, shrewd, and intelligent, Abe Kurosawa was seen as cunning, sly, and crafty.

This “damned if you do” phenomenon also produced the odd result of out-group leaders distancing their group from the group's own major accomplishments. Merton (1948) reported examples of Jewish leaders pointing out that Jews actually ran very few banks and of

expressing concern over the high proportion of Jewish doctors. Is this so odd? Merton (1948) put it this way: When Presbyterians take pains to point out that not that many Presbyterians are executives of major financial institutions, Jews doing so won't seem so unreasonable. And when the New York Yankees express concern about how often they win World Series, Jews expressing concern over the number of Jewish doctors will also seem reasonable.

This latter damned if you do phenomenon is interesting, and I agree that Merton (1948) has captured some truth to the nature of intergroup relations. However, as far as I can tell, this does not represent a self-fulfilling prophecy. In-group prejudice may influence in-group *evaluations* of an out-group, but they do not seem to be having much influence on the group's actual behavior or outcomes in these examples. A self-fulfilling prophecy would occur if the belief that the Jews ran the banks actually led Jews to become bank executives. And a belief among Jews that there are too many Jewish doctors would be self-fulfilling if it led *fewer* Jews to become doctors.

In general, though, Merton (1948) provided a thoughtful, creative, and compelling analysis of the ways in which erroneous social beliefs can become transformed into social reality. This is, perhaps, the most profound potential influence of internal states (prejudices, expectations, beliefs, etc.) on social perception because it involves fundamentally altering that which is perceived.

The Early Work: Some Preliminary Conclusions

The work on social perception was important and perhaps even revolutionary, because it suggested that the perceivers' own motivations and expectations helped actively construct social reality via two different routes. One was by influencing the perception without directly influencing the reality. That is, both the early stereotype work and the New Look research involved attempts to show that internal states (beliefs, stereotypes, prejudice, motives) influence the perception of external stimuli. In LaPiere's (1936) work, for example, the hospital manager's bigotry may have led him to see Armenians as making greater demands for public charity than they really made—but they did not cause Armenians to seek greater charity.

For a number of reasons, however, the early work on stereotypes and on the New Look was suggestive more than conclusive. Katz and Braly (1933) had no criteria with which to assess the degree of bias in Princeton students' perceptions of various racial and ethnic groups; LaPiere (1936) had good criteria, but actually provided only sparse evidence regarding the beliefs most non-Armenians held about Armenians; and F. Allport's (1955) critical review of the New Look pointed out that few, if any, of the New Look empirical studies provided clear evidence of internal states actually influencing perception. The one study from this time (Hastorf & Cantril, 1954) that provided both a clear assessment of beliefs and reasonable criteria for assessing bias and accuracy seems to have found far more evidence of accuracy than of bias.

Merton's (1948) proposal, however, was considerably more radical than those of the early social psychologists. Social beliefs not only biased perceptions of social stimuli (although he clearly included this phenomenon in his analysis)—they could also alter actual social reality. Although Merton (1948) told a good story based on historical and social patterns, he did not provide any original empirical data demonstrating the occurrence of self-fulfilling prophecies.

Indeed, his analysis lay more or less dormant for nearly two decades. And then it exploded into not only the scientific journals but also the covers of the mainstream press. How that happened is presented in the next chapter.

Notes

1. I do not mean this to be interpreted as suggesting that LaPiere was incompetent or that his study was hopelessly flawed. The study was creative, original, and groundbreaking when it was conducted in the 1930s. In some ways, it is unfair to criticize his study's quality by modern standards of social scientific research, which, of course, were unknown at the time. That being said, however, it is reasonable to evaluate the meaning and interpretation of his study, here and now, based on modern standards. It might have been reasonable and the best they could do for the ancients to conclude that the earth was the center of the universe. That does not make that conclusion justified. In the same manner, LaPiere's study was reasonable and innovative at the time. This does not necessarily mean, given what we know now about accuracy, bias, and methodology, that his conclusions were justified.

2. "Dominated by bias" could be taken to mean something much more strong: nearly 100% bias, at least twice as much bias as accuracy, etc. *At minimum*, however, it means more bias than accuracy. Therefore, treating "dominated by" this way makes it much easier to confirm the prediction that social perception was "dominated by bias" than requiring "dominated by" to mean nearly 100%.

3 The Once Raging and Still Smoldering Pygmalion Controversy

SELF-FULFILLING PROPHECY is a potentially unnerving phenomenon, both because it can threaten our hold on reality and because it raises the possibility that we create many of the social realities that we take for granted as being inherently “true.” How? Consider several contexts.

First, if people evoke from others what they expect from them, the validity of nearly all scientific research with humans is potentially threatened. Why? Because confirmation of a researcher’s hypothesis (expectation) may represent self-fulfilling prophecy rather than any fundamentally true scientific principle. Maybe that new drug only seemed to work because the researchers conveyed their positive expectations to participants in a drug trial study, and those positive expectations led those participants to adopt healthier behaviors. Thus, the new drug-takers may have become healthier, not because of the drug, but because of self-fulfilling prophecy. Similarly, perhaps those interviews with African American adolescents found evidence of an “oppositional identity” (rejection of all things believed to reflect “White culture or values,” such as education, achievement, career goals, etc.), mainly because that was what the interviewers expected to find, and they evoked such evidence from their respondents.

Second, if people evoke from others what they expect from them, perhaps children deserve little credit for their good qualities, and little blame for their flaws. Perhaps parents’ and teachers’ expectations cause children to become what they become. Is Isabela struggling in school? Well, maybe the problem is not her intelligence, motivation, or home life, but her teachers’ low expectations for her. Is Rudolf a behavior problem? Maybe he got that way because his parents always expect him to be a troublemaker. Saba may be valedictorian and captain of the high school soccer team, but she does not really deserve much personal credit for all that—it all may merely reflect others’ high expectations for her.

Third, if people evoke from others what they expect from them, perhaps broad inequalities between demographic groups, rather than reflecting characteristics of group members, or even the socio-cultural-historical context in which they find themselves, simply reflect the stereotype-based expectations of others around them. If Greek Americans are more likely than others to own diners, if African Americans are more likely than others to drop out of high school, and if women are more likely to decide to stay home to raise the kids, perhaps they are simply fulfilling others' stereotype-based expectations for them.

Self-fulfilling prophecy is potentially nasty stuff. It means that individuals cannot necessarily be held responsible for their failures; nor should they take too much credit for their successes. It also means that group differences in school achievement, income, occupation, alcoholism, criminal behavior, etc., may partially or even mostly reflect widely shared expectations for members of those groups. Although the early research did not directly address any of these potentially nasty implications of self-fulfilling prophecies, it did raise these issues. This chapter, therefore, reviews the early classic experiments demonstrating self-fulfilling prophecies and some of the theoretical and political controversies they raised.

Experimenter Effects: The First Empirical Studies of Self-Fulfilling Prophecies

Experimenter effects occur when researchers influence the outcomes of their studies in a manner that confirms their hypotheses. Research on experimenter effects provided the first empirical demonstrations of self-fulfilling prophecies. The early research on experimenter effects, therefore, is reviewed next.

The story of Rosenthal's dissertation. Despite the social sciences' long-standing interest in stereotypes, and despite Merton's (1948) interesting and compelling analysis, there were no systematic empirical investigations of self-fulfilling prophecies until Rosenthal's groundbreaking work on experimenter effects. As Rosenthal (1985) himself tells it, he stumbled onto the phenomenon while working on his dissertation. While intending to address the psychoanalytic concept of projection, he subjected college students to success or failure on an IQ-like test. Both before and after the test, he had participants rate the likely success or failure of individuals pictured in photographs. His projection hypothesis was that those receiving success feedback regarding their own IQ scores would show a substantial increase in their ratings of the successfulness of the folks in the photos; those receiving failure feedback regarding their own IQ scores would show substantial decreases in their ratings of the successfulness of the folks in the photos. And that is what he found.

So far so good. But here is the kicker. In a flash of intuition that something was not quite right, Rosenthal checked the *pretest* ratings—those obtained *before* the success/failure manipulation. He found that the group destined to receive the “success” feedback gave the *lowest* ratings—this group viewed the targets in the photos as *least* successful. There is no good reason for these ratings to have differed *before* people experienced success or failure. Why did this happen? Rosenthal suspected that he had unintentionally influenced the results. If so, this would not only be an “experimenter effect,” but would also be a self-fulfilling prophecy.

Apparently, Rosenthal unintentionally biased the results in a manner likely to confirm his projection hypothesis. How? Let's assume that this particular version of the projection

hypothesis is false, and that the success/failure manipulation had little or no effect on subjects. In the worst-case scenario, there would be no difference in the posttest ratings of the photos. *But if that happened, Rosenthal's success-condition subjects would show the largest increase in success ratings, thereby confirming his hypothesis.*

Consider Table 3-1. The posttest results clearly show no difference in ratings after experiencing success or failure oneself. Nonetheless, the *change* scores appear to support the projection hypothesis. Success appears to have led to a greater increase in success ratings than did failure. However, this is not because of anything having to do with the manipulation—it is only because the experimenters depressed the success ratings of the subjects during the *pretest*.

Rosenthal, being duly concerned about all this, went to members of the faculty, who replied, “Oh yes, we lose a few PhD dissertations now and then because of problems like that.” He then conducted an intensive review of the literature for research and writing about this type of experimenter bias. He uncovered quite a lot of speculative and anecdotal evidence suggesting that such effects were indeed fairly common, but there had not yet been any systematic research into the tendency for experimenters to find what they were looking for.

Experiments on experimenter bias I: Human subjects. I suspect that many young researchers, after running into a problem like this, would throw up their hands in frustration and turn to other endeavors. Not Rosenthal. He took this as inspiration for a program of research on experimenter effects that was to become the first systematic study of self-fulfilling prophecies.

First, Rosenthal and Fode (1963a) set out to see if they could systematically replicate Rosenthal's dissertation experience. They employed 10 experimenters—graduate and undergraduate students in a class on experimental psychology. The experimenters' job was to have students rate photographs, as in Rosenthal's dissertation. Half the experimenters were told to expect high ratings (+5, meaning success); half were told to expect low ratings (−5, meaning failure). The experimenters had about 200 students rate the photographs. The self-fulfilling

TABLE 3-1
Low Success Ratings on the Pretest, in the High-Success Condition, Bias the Study in Favor of the Projection Hypothesis

	Success	Failure
Pretest	3	7
Posttest	7	7
Change	4	0

Notes. All data are hypothetical. Higher scores indicate higher ratings of the success of targets in photos.

Results show that:

1. In the posttest scores, there is no difference in success/failure ratings after experiencing success or failure oneself (the manipulation).
2. The *increase* in success ratings in the success condition exceeds the increase in success ratings in the failure condition.
3. Thus, by artificially depressing the success ratings in the pretest condition, an experimenter can bias the study in the direction of support for the projection hypothesis (people see more success in others after they themselves have succeeded).

prophecy hypothesis is that the experimenters expecting research participants to provide success ratings would obtain higher success ratings than would experimenters expecting research participants to provide failure ratings. This is precisely what happened; and this pattern was replicated in two subsequent experiments.

Experiments on experimenter bias II: Animal subjects. At the time, behaviorism was still in its heyday, and many researchers worked with animals, especially rats. As Rosenthal (1985) tells it, when he discussed these findings with some of his colleagues, they replied, “Well of course you’d find expectancy effects and other artifacts when you work with human subjects; that’s why we work with rats.” So Rosenthal (Rosenthal & Fode, 1963b; Rosenthal & Lawson, 1964) set out to discover whether expectancy effects operated in research on rats.

In this research, experimenters (students in a lab course) were led to believe that rats had been bred to be either “bright” or “dull.” Brightness and dullness referred to their ability either to learn how to run a maze to obtain food (Rosenthal & Fode, 1963b) or to obtain food in a Skinner box (Rosenthal & Lawson, 1964). Half the experimenters were told that they would be working with bright rats; half were told that they would be working with dull rats. In fact, the rats had not been bred for brightness or dullness; they were randomly assigned to experimenters. Experimenters then spent several days training them.

The experimenter effect/self-fulfilling prophecy hypothesis is clear: The supposedly “bright” rats should learn to run the mazes more quickly than the supposedly “dull” rats. This is precisely what happened. In both studies, the “bright” rats outperformed the “dull” rats. Some of the results were downright amusing. In Rosenthal and Fode’s (1963b) study, nearly 3 in 10 “dull” rats did not even move from the starting point, whereas only 1 in 9 “bright” rats did not move. In both studies, experimenters admitted to being a lot nicer to the “bright” rats than to the “dull” rats. “Bright” rat experimenters described themselves as more relaxed, pleasant, friendly, and enthusiastic when working with the rats than did “dull” rat experimenters. And “bright” rat experimenters claimed to have handled their rats more frequently and more gently than did “dull” rat experimenters.

Experimenter effects were thus demonstrated to occur with both animal and human subjects. I was only in grade school when all this was going on, but I suspect that this work must have been fairly controversial. “Experimental psychology” (in general, behaviorism) dominated psychology departments around the country, and even within social psychology, the experiment had come to dominate empirical research. Then came Rosenthal (a new kid on the block) telling psychologists that their hypotheses may have been biasing their results all along. Not surprisingly, Rosenthal had difficulty publishing this early work in mainstream psychology journals (Rosenthal, 1985).

Thus, from its outset, empirical work on self-fulfilling prophecies appears to have been controversial. This was nothing, however, compared to the furor that erupted after Rosenthal turned his attention to elementary school teachers and students.

Teacher Expectations: The First Pygmalion Study

Despite the methodological importance of the work on experimenter effects, it was Rosenthal and Jacobson’s (1968a) classic and controversial Pygmalion study that launched the self-fulfilling prophecy as a major area of inquiry in the social sciences. After reading about

Rosenthal's research on experimenter effects, Lenore Jacobson, an elementary school principal, wrote to Rosenthal about her interest in teacher expectations, and suggested that, "If you ever 'graduate' to classroom children, please let me know whether I can be of assistance" (Rosenthal, 1985, p. 44).

By the early 1960s, the civil rights movement was in full swing, which helped sensitize many people to the role of racism in creating inequalities. The idea that teacher expectations could profoundly influence student achievement fit well with this social and political zeitgeist. Racism, as manifested in teachers' low expectations for, and unfair treatment of, minority students, could be a powerful contributor to educational inequalities. In fact, prior to the existence of any empirical research on the topic, there was fairly widespread belief, at least in educational circles, in the power of teacher expectations to create self-fulfilling prophecies (Rosenthal, 1985; Wineburg, 1987). In the absence of such research, at least one municipal policymaker cited Rosenthal's experimenter effects research on "maze bright" and "maze dull" rats in support of claims emphasizing the important role of teacher expectations in student achievement (see Wineburg, 1987, for a review).

Into this milieu stepped Rosenthal and Jacobson (1968a, 1968b), with their simple and ingenious Pygmalion study. They titled the book reporting this research *Pygmalion in the Classroom*, after a Greek myth in which Pygmalion, a sculptor, falls in love with his own statue—and his love and admiration for his statue is so intense that it actually brings the statue to life. Rosenthal and Jacobson administered Flanagan's Test of General Ability (the TOGA—a group-administered nonverbal intelligence test) to all of the children in Jacobson's elementary school (kindergarten through fifth grade) in the spring of 1964. However, they did not tell the teachers that this was an intelligence test. Instead, special covers conveyed that it was the "Test of Inflected Acquisition," which, an information sheet explained, was a new test being developed at Harvard for identifying children likely to "bloom"—to show a sudden and dramatic intellectual spurt over the upcoming school year.

Rosenthal and Jacobson then informed each teacher of the names of each of their potential "late bloomers." In fact, however, these students (about 20% of the total in the school) were selected at random. As Rosenthal and Jacobson (1968a, p. 70) put it, "The difference between the children earmarked for intellectual growth and the undesignated control children was in the mind of the teacher." They then administered the TOGA again 1 year later and 2 years later.

What happened next: The oversimplified version. Teacher expectations created a self-fulfilling prophecy. One year later the "late bloomers" gained more IQ points than did the control students. Even 2 years later, the bloomers' gains were still more than those of the control students. Although the only *initial* systematic difference between those and other kids was in the teachers' minds, the late bloomers actually became smarter—or at least their IQ test scores increased more (whether this actually represented students becoming smarter is discussed in more detail later in this chapter). The teachers' false belief had become a reality.

Teachers' expectations also colored their *impressions* of students. Rosenthal and Jacobson's (1968a, 1968b) teachers viewed the late bloomers as smarter—more curious, more interesting, and more likely to be successful later in life. As Chapter 2 pointed out, it had long been known that perceivers' expectations color their interpretations of targets' behavior, so this was not that surprising. At least somewhat more surprising was the finding that teachers also viewed the late bloomers as more pleasant—happier, more appealing, better adjusted, more affectionate, and less in need of others' approval.

One possible explanation for this pattern is that teachers were just reacting to genuine changes in the students' behavior and achievement. That is, because these kids really did become smarter and more successful in school, perhaps the teachers came to view them that way and liked them for it. If teachers were just reacting to positive intellectual changes in their students, one would expect teachers to have similarly positive reactions to (the smaller number of) kids in the control group who also showed dramatic IQ gains. This was not the case. Perhaps most surprising of all, Rosenthal and Jacobson's (1968a, 1968b) results showed that the more the *control* children gained in IQ, the *less* well adjusted, interesting, and affectionate they were seen by their teachers. In short, teachers seemed actively hostile toward the kids showing unexpected intellectual growth.

Taken together, and when described in this manner, these results seemed quite dramatic. Inaccurate high teacher expectations provided an undue advantage to some students, and inaccurately low expectations (i.e., failing to anticipate dramatic intellectual growth that did occur) seemed to trigger oppressive teacher responses. This was a powerful combination—and seemed to explain how teachers' expectations (and, by extension, expectations of managers, college admissions personnel, health professionals, etc.) could be a major contributor to social inequalities associated with race, sex, and social class.

What happened next: The messier, more complicated, and truer version. There is nothing false in the above, oversimplified version. It is true, and scientists have described the study in this manner for decades (e.g., Darley & Fazio, 1980; Fiske & Taylor, 1991; Gilbert, 1995; Myers, 1999; Schultz & Oskamp, 2000). Nonetheless, Rosenthal and Jacobson's (1968a, 1968b) pattern of results was not quite as straightforward as the oversimplified gloss seems to suggest.

First, let's review the major results regarding IQ change in more detail. One complication was that, on average, both groups of children—late bloomers *and* controls—showed dramatic IQ gains over the next year. On average, the late bloomers gained about 12 points and the controls about 8 points.

This is important for at least two reasons. First, in this study, there was *no IQ evidence of teachers "oppressing" (harming, stigmatizing, discriminating against, etc.) students*. Most students gained in IQ, regardless of condition. And the control group's average gain of 8 points is quite dramatic—it is about half of a standard deviation on a typical IQ test. Although the study's results did not preclude the possibility of teacher expectations actively harming students, there was not a shred of IQ evidence in this study indicating that such harm actually occurred.

Second, although the across-the-board IQ increases could be described as "dramatic," the *differences* between the gains of the late bloomers and the controls were not so dramatic. Averaging across all grade levels, that *difference* was about 4 points. This difference was statistically significant—but I think it would be difficult to characterize a 4-IQ-point difference as a "dramatic" effect. For example, if my daughter has an IQ of 120, and your daughter has an IQ of 116, I do not think you would consider my daughter to be dramatically smarter than your daughter, or even to have scored dramatically higher.

Other ways to consider the size of the effect also yield a picture of a less than dramatic result. The difference between experimental and control conditions corresponded to an effect size of 0.30 (difference between the experimental and control group in standard deviation units). Typically, effect sizes of 0.30 or less are not considered to be large (Cohen, 1988). Or, we could simply correlate the manipulation with IQ scores. That correlation is .15

(Rosenthal, 1985). The *size* of the difference between bloomers and controls in the Pygmalion study was something less than dramatic.¹

There was, however, *some* evidence of dramatic effects. In first grade, the bloomers out-gained the control kids by about 15 IQ points; in second grade the difference was about 10 points. In both grades, the control students gained IQ points—but such gains were not even close to those gained by the bloomers.

But the story again becomes complicated. There was no difference between third grade bloomers and controls. In fourth grade, bloomers gained more than controls, but the difference was not statistically significant. In fifth and sixth grade, bloomers actually gained *fewer* IQ points than did controls, but this difference was not statistically significant either.

So, the story was not so neat and clean. Still, one could, without too much difficulty, clean it up and tell a nice story. Maybe teacher expectations were not universally powerful influences on students' IQ. Maybe they did not always, or even usually, lead to self-fulfilling prophecies. Maybe it was just that very young children are far more susceptible to all sorts of adult influences (including teacher expectations) than are older children. Thus, one gets the large effects in first and second grade and virtually no differences thereafter.

This explanation, however, only *seems* to clean up the mass of complex and seemingly contradictory findings. Why? Remember that Rosenthal & Jacobson (1968a, 1968b) found that statistically significant differences between bloomers and control students persisted for 2 years. However:

1. The extent of those differences actually declined (from about a 4-point IQ difference between bloomers and controls to a less than 3-point difference); and
2. After 2 years, the *oldest* children showed the largest differences between bloomers and controls. So if there was much greater “susceptibility” among younger children, it did not last very long! And what mechanism could explain why, among the older children, there was a complete absence of a teacher expectation effect in Year 1 and the largest effects obtained in Year 2? This question has never been addressed by subsequent research and, frankly, I cannot even concoct a plausible explanation.

The Extreme Reactions to This Study

As we shall soon see, this type of inconsistency in the results provided ample opportunity to attempt to discredit the study. Nonetheless, the major pattern of results, especially when interpreted in the “oversimplified” manner, seemed both dramatic and to have profound implications for social problems and social policy. As a result, the study often evoked extreme reactions—usually positive from the general intellectual public and very negative among some educational psychologists. Because the study is such a classic, is so highly cited, and inspired several decades of research on self-fulfilling prophecies, I next discuss in some detail the extreme reactions (both positive and negative) to the study.

Uncritical acceptance of the study by the general intellectual public. The study hit a sensitive social and political nerve. It was published in the late 1960s, when liberalism was at a political peak, shortly after the passage of dramatic civil rights legislation. The consciousness of much of the country had been raised regarding the extent to which racism and discrimination

contributed to the massive inequalities between Whites and minorities. Left-wing intellectuals agreed that legal, institutional, and informal racism pervaded society and most of its White members. If we could do away with racists and racism, perhaps we could live together in harmony.

So then the Rosenthal and Jacobson (1968a, 1968b) study came along, and to this day, it has frequently been interpreted as demonstrating a widely generalizable mechanism of racial and social oppression (Coles, 1969; Gilbert, 1995; Hofer, 1994; Jones, 1990; Rist, 1970; Taylor, 1992; see Wineburg, 1987, for a review). Students come to achieve at levels their teachers expect of them. And, of course, because most teachers are White and middle class, they expect the most from White, middle class students and the least from non-White and poor students.

More generally, it has been cited in support of arguments claiming that, because teacher expectations are based heavily on social stereotypes, they are potentially a powerful force in the creation of social inequalities and injustices. Especially if this process occurs, not only in elementary school classrooms, but also in colleges, in the workplace, in government, etc., this phenomenon is capable of accounting for long-term maintenance of social inequalities. Thus, I suspect that Rosenthal and Jacobson (1968a, 1968b) so captured the imagination of the intellectual public, at least in part because its message was clear and simple and it seemed to provide scientific credibility and strong rhetorical ammunition for pundits, well-meaning policymakers, social activists, and reformers.

Later in this book, I address the extent to which self-fulfilling prophecies in general, and teacher expectations in particular, contribute to social problems and inequalities. For now, I will simply point out that (1) Rosenthal and Jacobson (1968a, 1968b) did not examine the role of stereotypes in the formation of teachers' expectations; (2) they only manipulated *positive* expectations, leaving as an open, empirical question the effects of negative expectations; (3) the effects they found were not particularly powerful; and (4) those effects became weaker over time. Viewed in this light, the study does not appear to provide much *terra firma* for claims regarding the power of teachers' expectations to create social injustices.

A 1994 *New York Times* Op Ed piece provides a classic example of the extent to which the Pygmalion study has been exaggerated and distorted in support of fundamentally political arguments. In the piece, Myron A. Hofer (identified by the *Times* as a professor of psychiatry at Columbia and director of a department of developmental psychobiology at a psychiatric institute) attempts to discredit the scientific validity of claims made in *The Bell Curve* (Herrnstein & Murray, 1994) regarding the role of genes in intellectual achievement. As part of his argument, he upbraids Herrnstein and Murray for failing to acknowledge or account for research findings that do not support their argument. And his prime case is Rosenthal's research:

In a typical experiment, elementary school teachers were told that 20 percent of their students had been found to be gifted through special tests. Actually, the names had been selected at random. By the end of the year, these children had gained an average of 15 points in IQ, while their classmates' IQ scores remained unchanged. Mexican-American children in the sample showed the greatest gains. (Hofer, 1994, p. A39)

One can only shake one's head and wonder at this description. Hofer never actually cites Rosenthal and Jacobson (1968a, 1968b)—he only refers to “highly relevant work by Robert

Rosenthal” and then describes this “typical experiment.” The first two sentences, however, aptly describe Rosenthal and Jacobson (1968a, 1968b) and I know of no other studies that ever produced a 15-point difference and included Mexican American children. The problems occur in the last two sentences in the quote. Neither Rosenthal and Jacobson nor any experiment of which I am aware can be adequately summarized as producing a 15-IQ-point difference between high-expectancy children and controls (there was a 15-point difference in first grade, but only a 4-point difference overall). Another error is the claim that the control children’s IQ scores did not change (they increased by 8 points). In addition, there was *no* IQ advantage for “blooming” Mexican American students (although they did have a small advantage in school grades). This type of description of the Pygmalion study typifies the ways in which its findings have been accepted uncritically, and then exaggerated and misrepresented in the general intellectual press, ever since the study’s publication.

The storm of criticism. Some researchers in educational psychology and intelligence went ballistic after evaluating the study (e.g., Jensen, 1969; Thorndike, 1968). Two (Elashoff & Snow, 1971) wrote an entire book critiquing the Pygmalion study. Consider the following, from Snow’s (1969) critique of Pygmalion that appeared in *Contemporary Psychology* (p. 197):

... The study suffers from serious measurement problems and inadequate data analysis. Its reporting, furthermore, appears to violate the spirit of Rosenthal’s own earlier admonitions to experimenters and stands as a casebook example of many of Darrell Huff’s (*How to Lie with Statistics*. New York: Norton, 1954) admonitions to data analysts.

Amusingly (at least to me), it seems that many of the complaints leveled against the original Rosenthal and Jacobson (1968a, 1968b) study were more flawed than the study itself. Although a rehashing of all the arguments for and against the paper is beyond the scope of this chapter (see, e.g., Elashoff & Snow, 1971; Rosenthal, 1974, 1985, 1995; Elashoff & Snow, 1971, 1995; Thorndike, 1968), I will briefly discuss some of the most flawed charges against the study.

One such charge was that the measure of IQ was unreliable (e.g., Roth, 1995; Thorndike, 1968) apparently in an attempt to suggest that any results developed using such a measure were meaningless. In fact, however, lack of reliability in a measure makes it *harder* to find differences between groups on that measure. Therefore, finding differences between groups with a measure low in reliability attests to the power of those differences.

Two other early criticisms were that (1) the IQ test used was not appropriately normed for the youngest children and (2) the scores of the children tested in kindergarten were so low (mean of 58) as to be manifestly invalid. There is some truth to both claims. However, the low scores probably occurred precisely because the test was not created for use on younger children. Furthermore, this critique is irrelevant to understanding Rosenthal and Jacobson’s (1968a, 1968b) results because it cannot, by itself, explain why they obtained a pattern of significantly greater IQ increases among the high-expectancy students. That is, even if the test was not created for use with younger children, how could such a test yield systematically and significantly higher IQ gains for the high-expectancy kids?

As late as 1984, when I was a graduate student beginning to study self-fulfilling prophecies, I approached a famous social psychologist for advice on how to pursue the topic. I was told, “There is no such phenomenon.” I said, “Why do you say that?” And the reply was, “The Rosenthal and Jacobson study was so flawed that it cannot be believed.” (Even then, I knew enough of the history of the phenomenon to not be deterred by this comment).

Taking Rosenthal and Jacobson’s (1968a, 1968b) Research at Face Value

A later section of this chapter addresses in detail some of the strongest evidence questioning the study’s validity. However, even if one takes Rosenthal and Jacobson’s results entirely at face value, the justifiable conclusions are more modest than suggested by the overly dramatic manner in which the study has frequently been portrayed.

This section is organized around several questions—questions to which *wrong* answers have often seemed obvious, or at least implied, in many discussions of the original Pygmalion study.

1. Were teacher expectations typically inaccurate? This was not assessed. Therefore, their study provided no information about the typical accuracy or inaccuracy of teacher expectations.
2. Did demographic-based stereotypes unduly bias expectations and perceptions? Rosenthal and Jacobson (1968a, 1968b) did not assess the extent to which student demographics or social stereotypes influenced teacher expectations. Therefore, the study provided no data bearing on the issue of whether stereotypes bias teacher expectations.
3. Were self-fulfilling prophecies typically powerful and pervasive? They were clearly not typically powerful. The overall effect size equaled a correlation of 0.15. The mean difference in IQ gain scores between late bloomers and controls was 4 points. Nor were they pervasive. Significant teacher expectation effects only occurred in two of six grades in Year 1 and in one of five grades in Year 2. Self-fulfilling prophecies did not occur in 8 of 11 grades examined.
4. Were powerful expectancy effects ever found? Yes. The results in first and second grade in Year 1 (15- and 10-point bloomer–control differences) were quite large.
5. Were self-fulfilling prophecies harmful? Rosenthal and Jacobson (1968a, 1968b) only manipulated positive expectations. They showed that false positive expectations could be self-fulfilling. It would have been unethical to instill false negative expectations. Therefore, they did not assess whether false negative expectations undermine student IQ or achievement.

The Immediate Follow-Ups

Because of the controversy surrounding Pygmalion, the first order of business for the scientific and educational community was to figure out whether the phenomenon was real. For the next 6 or 7 years, attempted replications were performed at an almost frantic pace,

both by Rosenthal and his colleagues, and by others (see reviews by Brophy & Good, 1974; Rosenthal, 1974). Even these studies, however, initially evoked considerable controversy. Consistently, only slightly over one-third demonstrated a statistically significant expectancy effect; almost two-thirds failed (Rosenthal & Rubin, 1978). This pattern seemed to resolve nothing and only add fuel to the fire. It was often interpreted by the critics as demonstrating that the phenomenon did not exist because support was unreliable. It was interpreted by proponents as demonstrating the existence of self-fulfilling prophecies because, if only chance differences were occurring, replications would only succeed about 5% of the time.

Resolution to the Furor

Rosenthal and Rubin's (1978) meta-analysis. Today, arguments about the strengths and flaws of the original Rosenthal and Jacobson (1968a, 1968b) study have become moot (mostly—but see the last section of this chapter). There have been hundreds of follow-up studies of the effects of expectancies in classrooms, the workplace, and laboratories. The Pygmalion controversy, however, was to have an effect that went well beyond self-fulfilling prophecies. In his attempt to refute critics, Rosenthal became one of the leaders in development of meta-analysis (Harris, 1991)—a statistical technique for summarizing the results of multiple studies. Although meta-analysis, too, was greeted with considerable skepticism by Rosenthal's critics (see the commentaries on Rosenthal and Rubin's [1978] meta-analysis in *Behavioral and Brain Sciences*), it has subsequently become the dominant method within the social sciences for summarizing the results of large research literatures and resolving controversies about the existence and size of effects.

Rosenthal and Rubin's (1978) meta-analysis of the first 345 experiments on interpersonal expectancy effects conclusively demonstrated the existence of self-fulfilling prophecies. The 345 studies were divided into eight categories. *Z* scores representing the combined expectancy effect for all studies in each category were computed. The median of the eight combined *Z* scores was 6.62. The likelihood of finding such a high *Z*-score by random chance alone is essentially zero, so that this provided conclusive statistical evidence that self-fulfilling prophecies occur.²

The naturalistic studies. One of the criticisms leveled against Rosenthal and Jacobson (1968a, 1968b) in particular, and experimental studies of expectancies in general, is that researchers induce perceivers (teachers, employers, etc.) to adopt false expectations by misleading or lying to them. For example, Rosenthal and Jacobson (1968a, 1968b) induced teachers to develop false positive expectations by claiming that certain students had been identified as late bloomers by a new test when, in fact, there was no test of late blooming. Therefore, such studies did not and could not address the extent to which teachers typically develop inaccurate expectations. This is important because only inaccurate expectations can produce self-fulfilling prophecies.

Naturalistic studies eliminated this problem by studying relations between naturally occurring teacher expectations and student achievement. Regardless of whether these studies used quasiexperimental techniques or survey/path analytic techniques, they consistently replicated Rosenthal and Jacobson's (1968a, 1968b) original finding that teacher expectations do indeed create self-fulfilling prophecies—usually with effect sizes closely corresponding to

those of that study (see Jussim & Eccles, 1995, for a review). So, over the last 30 years, whether Rosenthal and Jacobson (1968a, 1968b) is a good or bad study and did or did not find self-fulfilling prophecies has become a moot question. Hundreds of studies, both naturalistic and experimental, conducted in classrooms, laboratories, and a wide variety of other real-world contexts, have clearly shown that the self-fulfilling prophecy is a real phenomenon (see, e.g., the rest of this book, or reviews by Brophy, 1983; Brophy & Good, 1974; Cooper, 1979; Jussim, 1986; Rosenthal, 1974; Snyder, 1984).

Still Controversial After All These Years: Is the Effect on IQ Real?

You may now be thinking, “After a decade of heated debate, how could they have any energy left to argue?” Well, the controversy did die down, but at least a few combatants remained at the ready. But, you counter, “After the meta-analysis and all the naturalistic studies, what could they possibly have left to argue about? That self-fulfilling prophecies occur is now indisputable.” This is true. I know of no social science writing that denies the existence of self-fulfilling prophecies. What remains disputable to this day, however, is the viability of the most controversial claim made in the original Pygmalion study: that teacher expectations influence *IQ*.

IQ is not just any old dependent variable. IQ tests are intended to measure general intelligence—broad cognitive abilities for reasoning, planning, problem solving, learning, and thinking abstractly. IQ test scores often are the best predictors of many important life outcomes, including high school and college graduate rates; occupational success, income, and status; and likelihood of becoming an unwed mother or a convicted criminal (Detterman & Thompson, 1997; Herrnstein & Murray, 1994; Neisser et al., 1996). Intelligence clearly results from an interplay of genetic and nongenetic influences (e.g., Neisser et al., 1996). Nonetheless, it has been far easier for research to demonstrate a partial genetic basis for intelligence than to identify the environmental factors that lead to enduring changes in intelligence (e.g., Detterman & Thompson, 1997; Neisser et al., 1996).

In this context, the claim that teacher expectations influence IQ was extremely important, controversial, and difficult (for some) to believe. If 40 years of testing various experimental educational programs aimed at reducing disadvantage have failed to produce enduring increases in IQ scores (e.g., Detterman & Thompson, 1997), how likely is it that teacher expectations are endowed with such power? Few, if any, social scientists, educators, or educational psychologists deny the existence or importance of self-fulfilling prophecies in general. But there are quite a few who dispute the main claim of the original Pygmalion study—that teacher expectations influence student intelligence.

What is the rationale for disputing the effect on IQ? Different writers have made different arguments (see Spitz, 1999 for a review). Some (Roth, 1995; Rowe, 1995) simply repeated some of the classic criticisms of the original Pygmalion study (low reliability, invalid tests, etc.). If one had to rely exclusively on the Pygmalion study, some skepticism regarding the conclusion that teacher expectations influence IQ would be justified. It is, after all, only a single study and, like most single studies, has many important limitations. There have, however, been numerous follow-ups. Some Pygmalion detractors, however, have also addressed the subsequent research.

The saga of Wineburg and Raudenbush. In a paper titled “The Self-Fulfillment of the Self-Fulfilling Prophecy,” Wineburg (1987) provided one of the most sweeping assaults on the IQ effect. First, Wineburg provided a conceptual and historical review documenting how the social and political zeitgeist set the stage for popular acceptance of the study. Next, Wineburg summarized many of the early critiques of the Pygmalion study in a manner that strongly implied they invalidated Pygmalion’s conclusions, but without explaining how the supposed flaws could have led to the observed systematic differences between the expectancy and control conditions.

Nonetheless, even Wineburg (1987, p. 34) recognized the existence of self-fulfilling prophecies:

Within education, the issue had never been whether teachers form expectancies or whether these expectancies affect students. . . .

The bone of contention for Wineburg (1987, p. 34) was IQ:

Obscured and long-forgotten, the heart of the Pygmalion controversy was the bold claim that intelligence was affected by teacher expectations.

Wineburg (1987) then proceeded to review the follow-up studies that focused exclusively on *intelligence*. That review highlighted the weak to nonexistent effect often found on IQ of the follow-ups. Shortly before Wineburg published his paper, however, Raudenbush (1984) published a meta-analysis of the effect of experimentally induced teacher expectations on IQ. Wineburg (1987, p. 34) described that meta-analysis as follows:

In a meta-analysis based on 18 studies, Raudenbush (1984) found a small mean effect size in IQ expectation studies ($d = 0.11$), a finding that either achieved or failed to achieve statistical significance depending on the test employed.

Strictly speaking, there is nothing false here. An effect size of 0.11 is very small (corresponding to a correlation of about 0.06), and Raudenbush did indeed test for statistical significance in several ways, some of which showed that the effect was reliable and some of which did not.

But Wineburg (1987) either missed the main point of Raudenbush’s (1984) paper or chose to ignore it. One of Raudenbush’s (1984) main hypotheses was that the time of year that the study was conducted was a crucial moderator of expectancy effects. Why? Early in the year, teachers have had little direct experience with their students. In general, all they have is information from their records (previous grades, standardized test scores, etc.) and, perhaps, comments from other teachers. Consequently, they might find new information, such as that provided by a new test of “late blooming,” to be very useful indeed.

In contrast, the later the expectancy induction, the less likely it might be to actually change teachers’ expectations. By December, for example, teachers have had extensive contact with their students and have had the opportunity to see for themselves their performance on tests, on homework, and in class. Thus, they might be far more likely to discount the importance or validity of a test whose results seemed inconsistent with their direct experience with the student.

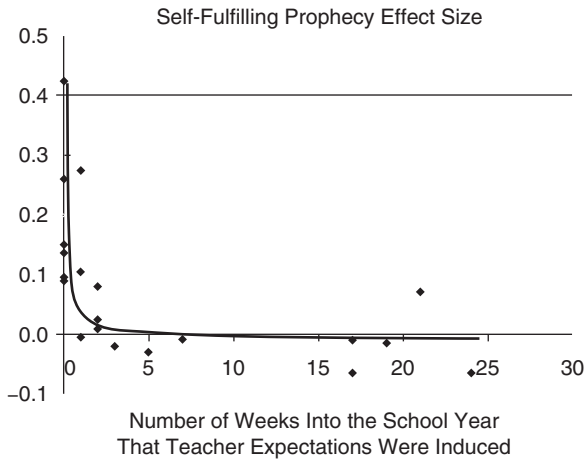


FIGURE 3-1 Relationship between time of year of the expectancy induction and the expectancy effect size. Adapted from Raudenbush (1984).

Determining this, rather than the overall effect size, was one of the main purposes of Raudenbush's (1984) study. And did he ever find this. Figure 3-1, adapted from his results, displays the relationship between time of year of the expectancy induction and the expectancy effect size. Figure 3-1 is so dramatic, it passes the most stringent test of all—the IOI (interocular impact test—i.e., the pattern hits you between the eyes). The relationship between time of year of induction and effect size is strongly *curvilinear*. Effect sizes closely corresponded to the Pygmalion effect of $r = 0.15$ when the manipulation was conducted within the first week of the year, but then rapidly dropped off after that. Expectancy inductions more than 2 weeks into the school year essentially produced no effect. In addition to the IOI, Raudenbush (1984) showed that this curvilinear relationship was highly statistically significant.

So what is the take-home message here? Raudenbush's (1984) meta-analysis found only a very small overall effect of expectancy manipulations on IQ. This was not, however, because such effects usually hovered near zero. More credible expectancy manipulations—that is, those conducted early in the year—were far more likely to produce expectancy effects on IQ. Even those strongest effects, however, were not particularly powerful, corresponding to a correlation of about 0.15.

The Later Exchanges

In 1994, Rosenthal updated the 1978 Rosenthal and Rubin meta-analysis with more recent research but reached essentially the same conclusions. In a reply, Snow (1995) emphasized that he agreed that self-fulfilling prophecies were a genuine and important phenomenon. However, he pointedly argued that there was no evidence supporting the notion that teacher expectations influence *intelligence*. He provided an intriguing reanalysis of the original Pygmalion data, which pointed out that many of the first and second graders' scores (those among whom the expectancy effect was strongest) were quite bizarre: Some students had

pretest IQ scores near zero, and others had posttest IQ scores over 200. Obviously, however, the children were neither vegetables nor geniuses. Furthermore, Snow (1995) pointed out that the TOGA's scores were only normed for scores between 60 and 160. If one excluded all scores outside this range, the expectancy "effect" disappeared.

Moreover, there were five "bloomers" with wild IQ score gains: 17–110, 18–122, 133–202, 111–208, and 113–211. Does anyone really think that the first two kids went from being vegetables to reasonably smart, or the last three went from reasonably smart to extraordinary genius, as a result of teacher expectations? If one simply excluded these five bizarre gains (outliers, they might be called by the statistically inclined), the difference between the bloomers and the controls evaporated.

Snow (1995) also attempted to discredit the conclusion reached in Raudenbush's (1984) meta-analysis (see Rosenthal, 1995, for a reply). In short, Snow (1995) highlighted the fact that some of the teacher expectation IQ studies produced reversals (higher IQ scores or gains in the control group) and argued that the minuscule median effect size of 0.035 was a better estimate of the effect than the mean effect size.

However, at about the same time, Raudenbush (1994) published a reanalysis of the 18 experiments included in his earlier meta-analysis using random effects models, which permit greater generalization than did his earlier method assuming fixed effects. The effect size for the four studies in which there was no prior teacher–student contact was 0.43, corresponding to a correlation of about 0.2 between expectancy manipulation and IQ.

Conclusion

Can any general conclusions be reached from what may appear to be a mess of complex findings, inconsistent replications, and heated controversy? I think so—and they are quite important. Erroneous teacher expectations, at least sometimes, create self-fulfilling prophecies. This issue is not in dispute—even the diehard critics of the original Pygmalion study agree that this is a fact (Snow, 1995; Spitz, 1999; Wineburg, 1987).

Self-fulfilling effects of teacher expectations are typically small. Although self-fulfilling prophecies in the classroom are real and occasionally large, far more often, they tend to be small. Although these conclusions are also old news in some circles, the periodic resurgence of claims emphasizing the power of self-fulfilling prophecies (Claire & Fiske, 1998; Jones, 1986; Jost & Kruglanski, 2002; Schultz & Oskamp, 2000) suggests that even this old news bears reaffirmation.

Caveats to Pygmalion. Although controversies surrounding Rosenthal and Jacobson's (1968a, 1968b) study have been well-known for years, this chapter has documented the frequency with which Pygmalion is still summarized in an uncritical, oversimplified manner that consistently distorts the results. The Pygmalion study has been used to justify arguments claiming that expectancy effects are powerful and pervasive, intelligence is primarily environmentally determined, and relatively simple interventions can improve student achievement. It has also been used to justify arguments emphasizing the power of beliefs to construct social reality. Such uses of Pygmalion are not restricted to claims published before 1973, or even before 1993. For the many researchers who may not be aware that the entire self-fulfilling prophecy effect hinged on the occurrence of bizarre outliers and out-of-range IQ scores, the

sections reviewing Snow's various critiques (Elashoff & Snow, 1971; Snow, 1969, 1995) documenting this state of affairs should constitute some eye-opening news.

Many social scientists, however, may be aware of these weaknesses but choose to ignore them when discussing Pygmalion. It is, of course, a matter of scientific judgment how much of any study to believe. Therefore, another important goal of this chapter has been to document the highly limited and constrained nature of the conclusions justified on the basis of the Rosenthal and Jacobson (1968a, 1968b) study, even if its results are taken entirely at face value.

Putting controversy in perspective. Another purpose of this chapter has been to point out that, although debate between the different positions is often heated, the degree of factual disagreement between them is actually quite small. If one believes the critics, the IQ effect is zero. If one believes the advocates, it is very small (frequently 0, never consistently much higher than an r of 0.2). This chapter has not resolved this remaining degree of disagreement. It has pointed out, however, something that may have been lost in the heat of the controversy: Although the scientific evidence may be equivocal regarding whether teacher expectation effects on IQ are nonexistent or reliably very small, it is completely unequivocal that such effects, if they occur at all, are not very large by any standard.

Nonetheless, I would argue that for many purposes, whether such effects also occur for intelligence (or, at least, for IQ test scores) may not be all that important. First, consider the issue from the standpoint of a parent of school-age children. For the sake of argument, I am willing to stipulate that teacher expectations have no effects on IQ—they “only” influence students’ grades, standardized test scores, motivation, and the quality of their interactions with their teachers. If my daughter has a teacher with inordinately low expectations for her, and if she ends up disliking school, learning less, and receiving lower grades and standardized test scores, I would be extremely upset—and I would not be the least bit assuaged to discover that her IQ scores remained unchanged. If such patterns continued through high school, they could have a profound influence on the quality of the college my daughter would be capable of entering. However, after I calmed down, I would realize that although the situation still stinks, an effect on intelligence could be much worse—had the teacher actually lowered her intelligence, it would be that much more difficult to overcome all the other obstacles the teacher imposed on her.

Second, consider the issue from the standpoint of social problems. Again, let’s stipulate that teacher expectations have no influence on IQ; therefore, they cannot possibly account for the existing IQ differences between racial/ethnic groups. However, the potential influence of teacher expectations on grades, motivation, and standardized test scores would seem important enough. Teacher expectations could account for at least part of the large differences between racial/ethnic groups on major standardized achievement tests, such as the SATs and GREs. They could account for at least part of the major performance differences between racial/ethnic groups in high school and college. And if teachers’ low expectations for minorities contribute to the creation of a hostile learning environment, many minority students may disengage from education altogether (e.g., Steele, 1992).

Whether teacher expectations influence intelligence remains an important issue for future research in this area. Regardless, it is clear that self-fulfilling prophecies do happen, and they influence a wide array of student outcomes. For many practical purposes, the potential effect of teacher expectations on so many major academic outcomes may be far more important than an absence of an effect on IQ test scores.

Despite the controversies and the real or imagined flaws of their study, Rosenthal and Jacobson (1968a, 1968b) (1) created a strong and clear methodology for studying expectancy effects (i.e., experimental induction of false expectations) and (2) inspired a great blossoming of research on expectancy effects. The story of that blossoming is told in the next section.

Notes ---

1. For the statistically uninitiated, an effect of 0.30 standard deviation is comparable to a 30-point difference on standardized tests such as the SATs or GREs. A correlation of 0.15 is also quite small. Correlations assess how strongly two variables—in this case, teacher expectations and student IQ test scores—go together. Correlations can range from -1.0 to $+1.0$; 0.15 can be interpreted to mean that teacher expectations substantially increased the IQ scores of about 8% of the children in the late bloomer conditions (Rosenthal, 1985). This, of course, is the same thing as saying that teacher expectations did not greatly increase the achievement of 92% of the students. Thus, the effect size was not particularly large or dramatic.

2. For the statistically disinclined, Z scores are a way to relate raw scores to probability, or to a percentile rank. A Z score of 0 is at the 50th percentile. A Z score of 1 is at about the 84th percentile. A Z score above 3 is in the 99th percentile. When computing effects for experiments, these are useful because they indicate the probability that luck alone could have caused a result. The Z score over 6 obtained in Rosenthal and Rubin's (1978) meta-analysis is so high that, for all practical purposes, it means the self-fulfilling prophecy effects obtained in research could not possibly have occurred by chance alone.

2 The Awesome Power of Expectations to Create Reality and Distort Perceptions

This page intentionally left blank

4 The Extraordinary Power of Self-Fulfilling Prophecies

The Case for Powerful and Pervasive Expectancy Effects: Introduction

Many social scientists were about to fall in love with expectancy effects. Such effects were seen almost everywhere and much scientific writing emphasized their power and pervasiveness. The attention received by the early classics described in this chapter, even today, often dwarfs that of the attention given to failures to replicate those same studies. Accuracy, which provides a strong alternative explanation to self-fulfilling prophecies for why people's expectations predict social reality, was not addressed in this research.

The two chapters in this section review the early self-fulfilling prophecy and expectancy-confirming bias classics in such a manner as to capture the spirit of the thinking about interpersonal expectancies that was once present in much scholarly writing. Table 4-1 identifies the variety of expectancy-related phenomena addressed in the next several chapters. Table 4-2 specifically identifies two separate places in this book where these expectancy-related phenomena are discussed. Each phenomenon is discussed at least twice: once in the upcoming unabashedly enthusiastic section on expectancy effects and again in the next section (Chapters 6 through 9), which critiques the early research and emphasizes limitations to expectancy effects.

One major disadvantage to organizing these next several chapters in this manner is that the critique of research on a particular phenomenon does not come till two or more chapters *after* my first presentation of that research. For example, the critique of the early research on self-fulfilling prophecies (Chapter 6) comes two chapters after my enthusiastic review of the early research (Chapter 4).

I have, however, decided to stick with this organization for several reasons. First, a presentation-followed-by-immediate-critique organization would create a very nuanced, complex,

TABLE 4-1

Types or Classes of Expectancy Effects	
Behavioral Effects	What Phenomenon Is This?
An originally false social belief leads to its own fulfillment.	Self-fulfilling prophecy
Behavioral confirmation (targets confirm erroneous stereotypes or erroneous teacher expectations).	Self-fulfilling prophecy
Perceivers' expectations influence their behavior toward targets. ^a	<i>Potential</i> ^b mediator of self-fulfilling prophecy
People seek information that confirms their expectations. ^a	Biased information-seeking. This, too, is a <i>potential</i> ^b mediator of self-fulfilling prophecies.
<i>Cognitive Effects</i>	
Perceivers' expectations bias their evaluations or judgments of targets.	Expectancy-confirming evaluative or judgmental biases ^c
Attributions for targets' behavior are biased by expectations.	Expectancy-confirming attributional biases ^c
People better remember expectancy-consistent target information than expectancy-inconsistent target information.	Expectancy-confirming memory biases ^c

^a Biased information-seeking could reasonably be considered a subset of “perceivers’ expectations influence their behavior toward targets.” In the chapters ahead, however, I consider it as a separate topic because the research literature on biased information-seeking has developed largely independently of the research literature addressing how perceivers act on their targets.

^b Perceiver behavior and information-seeking may reflect their expectations, but will only create a self-fulfilling prophecy if that behavior or information-seeking evokes expectancy-consistent behavior from targets.

^c Inasmuch as this book focuses on interpersonal expectations, I often drop the “expectancy-confirming” part and refer only to “evaluative biases,” “judgmental biases,” “attributional biases,” or “memory biases.”

and limited view of the power and pervasiveness of expectancy effects. Such an organization could not possibly capture the excitement created by the early research. Furthermore, I suspect that it would be extremely difficult for most readers unfamiliar with this area to then understand why so many of the early major reviews of expectancy effects (and some current ones, too) so strongly emphasized expectancy effects. By conveying a sense of this initial enthusiasm, I hope to provide some insight into the good reasons why so much writing about expectancy effects has emphasized their power and pervasiveness (indeed, I could not think of a better way to explain why this research is still commonly discussed or cited in a similarly uncritical and enthusiastic manner to this day—e.g., Jost & Kruglanski, 2002; Ross, Lepper, & Ward, 2010; Weinstein, Gregory, & Strambler, 2004—than to write a chapter from an enthusiastic and uncritical perspective).

Last, any reader who is not interested in understanding the sources of this enthusiasm is welcome to read the chapters out of chronological order. For example, a reader interested only in the early research on self-fulfilling prophecies might consider reading Chapters 4, 6, 7, and 8 together. A reader primarily interested in the early research on expectancy-confirming

TABLE 4-2

Which Phenomena Are Discussed in Which Chapters?		
Phenomenon	Enthusiastic Review of Early Research ^a Presented in:	Critique/Limitations of Early Research ^a Presented in:
Self-fulfilling prophecy	Chapter 4	Chapters 6, 7, 8
Perceiver expectations influence their own behavior, which mediates self- fulfilling prophecies	Chapter 4	Chapter 6, 8
Information-seeking bias	Chapter 5	Chapter 8
Evaluative/judgmental bias	Chapter 5	Chapter 9
Attributional bias	Chapter 5	Chapter 9
Memory biases	Chapter 5	Chapter 9

Note. Bias here refers to expectancy-induced bias. There are many types of biases not addressed in this book.

^a These chapters *are not* intended to provide a comprehensive review of all research on interpersonal expectancy effects. They focus exclusively on the early research. In this book, “early research” typically refers to research performed in the 1970s and 1980s. There are, however, two exceptions: (1) More recent studies (e.g., 1990 and later) that specifically attempted to replicate the early research is also discussed in these chapters, and (2) several meta-analyses addressing various expectancy effects were published in the late 1980s and early 1990s—because nearly all of the studies included in these meta-analyses were performed before 1990, these meta-analyses are also discussed.

Chapter 3 reviewed the early research on teacher expectations and self-fulfilling prophecies; Chapters 13 and 14 review the more modern research on teacher expectations. Chapters 16 through 19 review evidence on stereotypes, some of which also addresses bias. Chapter 20 reviews some recent self-fulfilling prophecy studies that did not neatly fit into earlier chapters.

biases might consider reading Chapters 5, 8, and 9 together.¹ Table 4-2 is included here in part to serve as a guide to anyone interested in skipping chapters in order to focus on a specific topic.

Nonetheless, I think it is important to understand the extraordinary extent of psychology’s love affair with expectancies. To do so, I next explore the explosion of research that followed on the heels of the classic and controversial Pygmalion study.

Social Psychology Falls in Love with Self-Fulfilling Prophecies

Although most social psychologists avoided the intellectual battles surrounding the Rosenthal and Jacobson (1968a, 1968b) study, Pygmalion also had a profound influence on social psychology. Rosenthal and Jacobson (1968a, 1968b) had raised the possibility that self-fulfilling prophecies were a widespread and common phenomenon. And perhaps even more important, they provided a basic methodology (induction of false expectations) for experimentally testing for self-fulfilling prophecies. Armed with this information, and with a long-standing theoretical and applied interest in stereotypes and prejudice, inspired by the evidence of the existence of self-fulfilling prophecies and by the beginnings of the “cognitive revolution” in psychology, the golden age of social psychological research on expectancy effects was about to begin.

This chapter does not review every individual study of self-fulfilling prophecies. In 1978, Rosenthal and Rubin published a review and meta-analysis titled “Interpersonal Expectancy Effects: The First 345 Studies.” I am not going to review 345 studies (more, actually, because of course lots of studies have been published since then). Instead, I have the following criteria for selecting a relatively small number of studies to review here: (1) In my view, they capture the spirit and are representative of much of the early research on self-fulfilling prophecies; (2) many are classics that are, in essence, “must-cites” and can be found in most subsequent reviews of self-fulfilling prophecies; and (3) some of the less well-known studies included here may actually provide clearer evidence of self-fulfilling prophecies than some of the well-known classics (i.e., in this chapter, my goal is to paint a picture as favorable as possible regarding the perspective that expectancy effects are powerful and pervasive).

This chapter does not, however, review any research on any type of expectancy effect other than a self-fulfilling prophecy (i.e., a change in a target’s² behavior or attributes). This means that all sorts of expectancy effects that *do not* involve changing a target’s behavior or attributes (e.g., effects of expectancies on perceivers’ judgments, attributions, memories, or information-seeking) are *not* addressed in this chapter (such effects are the subject of Chapter 5). That said, on to the early classic research on self-fulfilling prophecies!

Self-Fulfilling Stereotypes

For many social psychologists, the links of expectancy effects to stereotypes and prejudice were obvious, as were the implications of expectancy effects for understanding social injustices and inequalities (see, e.g., Darley & Fazio, 1980; Devine, 1995; Fiske & Taylor, 1984, 1991; Jones, 1990; Weinstein et al., 2004). The main ideas were straightforward³ and typically started with the following premises:

1. Stereotypes are people’s beliefs about groups and their individual members that are typically inaccurate, rigid, negative, irrational, and resistant to change (see Allport, 1954; Ashmore & Del Boca, 1981; Brigham, 1971; Jussim, McCauley & Lee, 1995, for reviews; note: this view is itself critically evaluated in Chapters 15 through 19, but, for now, it is sufficient to acknowledge that many researchers hold this view).
2. Stereotypes are often a major source of expectations regarding individuals from the stereotyped groups. If stereotypes are inaccurate, they would typically lead to inaccurate expectations.
3. Expectations are self-fulfilling.

This leads to the following hypothesis:

4. Stereotype-based expectations, by leading to self-fulfilling prophecies, will create differences between individuals from different social groups, even when no real differences between the groups exist. When evidence of such stereotype-based self-fulfilling prophecies had been found, the inexorable conclusion was:
5. Self-fulfilling prophecies could be a major contributor to social inequalities.

The actual empirical studies, many of which have become classics in social psychology, focused primarily on steps 2 through 4. Point 1 was simply taken as a given by most researchers until about the 1990s (when many researchers began empirically assessing the accuracy of stereotypes—see Chapters 15 through 19). Inasmuch as point 5 refers to broad societal patterns involving millions of people, it could not possibly be empirically demonstrated by the type of relatively small-scale experiments demonstrating self-fulfilling prophecies. But if stereotype-based expectancies were self-fulfilling in small-scale laboratory contexts, the reasoning went, they were probably even more powerful in daily life (e.g., Darley & Fazio, 1980; Fiske & Taylor, 1991; Snyder, 1984).

Racial Stereotypes

The racist interviewer studies. Word, Zanna, and Cooper (1974) performed the first experiments that examined the potentially self-fulfilling effects of a social stereotype—specifically, racial stereotypes. This landmark research examined whether Whites' stereotypes and prejudice regarding African Americans could undermine the competence and performance of African Americans. If so, then self-fulfilling prophecy, rather than genuine competence differences between Whites and African Americans, could account for the continued lower social status, lower income, and lower educational attainment of African Americans.

Their studies focused on job interview situations, in part because of the obvious relevance of this situation to occupational success and, by implication, race differences in occupational status and income. Perhaps stereotype-based self-fulfilling prophecies reduced African Americans' opportunities to obtain good jobs by undermining their performance in interviews. If they performed more poorly in interviews, then they would be less likely to be hired.

Word et al. (1974) performed two experiments. In the first, White perceivers interviewed targets for a job. In fact, however, targets were confederates who had been carefully trained to engage in the same set of behaviors with each interviewer. Half the confederate targets were African American and half were White. The main dependent variables were interviewers' nonverbal behavior. Consistent with a self-fulfilling prophecy, perceivers were colder to African American targets than to White targets. In comparison to White targets, interviewers sat farther away from African American targets, had more speech disfluencies when talking to them, and conducted a shorter interview.

In their second experiment, Word et al. (1974) showed that this treatment undermined the performance of interviewees. Confederates were trained to interview applicants in either of two ways: (1) the cold style comparable to that received by the African American interviewees in Study One or (2) the warm style comparable to that received by the White interviewees in Study Two. All subject-applicants in this study were White. Results showed that the applicants treated coldly, as were the African American applicants in Study One, actually performed more poorly in the interview than did the applicants treated warmly. In comparison to the warmly treated applicants, the applicants treated coldly made more speech errors and independent judges rated their performance more poorly. The type of treatment accorded African American applicants in Study One undermined the actual interview performance of White applicants in Study Two.

One of the great things about this second study is that all of the applicants were White. Therefore, whether or not there were any real differences between the quality of African Americans' and Whites' interviews is irrelevant. Even in the utter absence of preexisting differences between White and African American job candidates, Whites' stereotypes and prejudice may be sufficient to undermine African Americans' performance in the interview. If so, stereotype-based self-fulfilling prophecies would seem to represent a major obstacle to African Americans' employability.

Another major strength of their research was that they examined nearly every step in the self-fulfilling prophecy process. Perceivers developed different expectations for different targets; different expectations led to differential treatment of targets; and targets reacted to the differential treatment. Thus, they seemed to have mapped nearly the entire self-fulfilling process in interracial interaction.

Other suggestive research on racial stereotypes and self-fulfilling prophecies. There were no published attempts to replicate the research by Word et al. (1974) for over 20 years (the one partial replication—Chen & Bargh, 1997—will be discussed later in this book). Furthermore, I know of no research on potentially self-fulfilling racial stereotypes from this time period that even attempted to map the entire process, as did Word et al. (1974). Nonetheless, two studies conducted at about the same time did address whether teachers (actually, teachers in training) treated White students differently than they treated African American students.

A study by Rubovitz and Maehr (1973) examined whether college students in teacher training courses treated African American and White students differently. For the most part, they did. They provided considerably more attention, encouragement, and praise to the White students, and they were less likely to criticize them or ignore them. A similar study conducted a few years later found no overall tendency for college students in teacher training courses to treat White students more positively (Taylor, 1979). Instead, Taylor (1979) found a race \times sex interaction, such that teachers in training treated White males most positively and African American males *least* positively.

Neither Rubovitz and Maehr (1973) nor Taylor (1979) assessed whether the differential behavior of teachers in training toward African American and White students influenced those students' performance. Thus, neither demonstrated self-fulfilling effects of stereotypes or prejudice. This was not the intended goal of either study. Instead, in both studies, the researchers' explicit goal was to begin identifying some of the interpersonal processes that could mediate self-fulfilling prophecies in the classroom. Both found evidence that was consistent with the idea that teacher stereotypes and prejudice could damage the achievement of African American students. Thus, both found results that supported the burgeoning consensus among social psychologists and other social scientists that stereotype- and prejudice-based self-fulfilling prophecies could have profound influences on students' achievement.

Sex Stereotypes

The "fake males" study. Skrypnec and Snyder (1982) used an ingenious method for examining the potential self-fulfilling power of sex stereotypes. Unacquainted pairs of males and females interacted in different rooms, entirely by a system of flashing lights. That is, they neither saw nor spoke to one another. This was crucial, because all males were assigned to the perceiver

role; all females were targets. However, because they neither saw nor spoke to one another, Skrypnec and Snyder (1982) were able to manipulate the male perceivers' *beliefs* about whether their partner was male or female. After providing a cover story about studying "minimal communication" (to justify the separate rooms and light communication system), they presented the male perceivers with a questionnaire supposedly completed by their partner. In the "male label" condition, this questionnaire appeared to be completed by a 20-year-old male sophomore who (supposedly) described himself in a stereotypically masculine manner (independent, athletic, assertive, etc.); in the "female label" condition, this questionnaire appeared to be completed by a 20-year-old female sophomore who (supposedly) described herself in a stereotypically feminine manner (shy, gentle, unathletic, etc.). Thus, even though all targets were actually female, half the time, the male perceivers believed that they were male.

Their task was to divide up among themselves 24 tasks using the system of lights to communicate who would do what. The tasks ranged from highly "masculine" (bait a fishing hook, etc.) to neutral (code test results) to highly "feminine" (decorate a birthday cake). The main self-fulfilling prophecy hypothesis was that targets labeled as male would end up (after a few rounds of negotiating) agreeing to relatively more masculine/less feminine tasks than would the targets labeled as female.

The results confirmed this prediction. In addition, Skrypnec and Snyder (1982) found that perceivers acted very differently with partners they believed were male than with those they believed were female. Specifically, perceivers started off the task negotiation process selecting for themselves more masculine tasks when they thought their partners were female than when they thought their partners were male. Similarly, when their choices conflicted, perceivers were less likely to give in to their supposedly female partners than to their supposedly male partners.

Bringing target sex under experimental control in an actual social interaction is no easy task. Thus, the great strength of this study was the creative way the researchers manipulated the apparent sex of the targets, and then followed how perceivers' beliefs about targets' sex influenced the course of the social interaction in such a manner as to fulfill sex stereotypes.

The sexist interviewer studies. A series of studies by Zanna (Zanna & Pack, 1975; von Baeyer, Sherk, & Zanna, 1981) used a very different approach to examine the potential self-fulfilling power of sex stereotypes. Rather than manipulating perceivers' expectations, they manipulated targets' beliefs about the sex-role attitudes of perceivers. They then examined the conditions under which targets' behavior conformed to sex-role stereotypes.

The research participants, all of whom were female, were led to believe that they would (at a later time) interact with a male participant supposedly to judge the accuracy of each other's initial impressions. Those initial impressions were to be based on questionnaire responses. Thus, each female participant received questionnaire responses supposedly from the person they would interact with, and they provided responses on the same types of questions.

The questionnaire from the supposed male partner contained two experimental manipulations. First, it manipulated the males' attractiveness. He either described himself as a 6-foot-1-inch-tall 21-year-old Princeton senior without a girlfriend but interested in meeting female college students (the attractive condition), or as a 5-foot-5-inch-tall 18-year-old non-Princeton freshman with a girlfriend who was not interested in meeting other female college students (the unattractive condition). In addition, half the time he described his attitudes

toward women's roles as either very traditional (e.g., liking women to be home oriented and passive) or nontraditional (e.g., liking women to be independent and ambitious).

One main dependent variable was how the women presented themselves to their (supposed) male partner. This was assessed via the questionnaire that the women completed about themselves. The women believed that their responses to this questionnaire would be given to their male partners; thus, the women could alter those responses to appeal to their partner if they so chose. This is exactly what happened, but only when the male was supposedly attractive. In that condition, women presented themselves as subscribing to much more traditional sex-role attitudes when they believed the attractive male held traditional attitudes than when he held nontraditional attitudes. There was no difference in how the women presented their sex-role attitudes toward the unattractive male partner.

Score on an anagrams test (presented to the women as a quick intelligence assessment) was a second dependent variable. Again, consistent with the self-fulfilling prophecy hypothesis, the women successfully unscrambled *fewer* anagrams when they believed their partner held traditional (compared to nontraditional attitudes), but this difference occurred only in the attractive male partner conditions. There was no difference in their scores in the traditional and nontraditional conditions for the unattractive male partner.

One limitation to the Zanna and Pack (1975) study was that no interaction between interviewer and interviewee ever took place. This limitation was addressed in a follow-up study (von Baeyer et al., 1981). Women were led to believe they were participating in a study of job interview techniques, and that graduate students had to conduct real interviews as part of this study. They were led to believe the graduate student interviewers thought that the job (and, therefore, the interview) were real, and they were encouraged to take the interview seriously because it provided a good chance to practice real job interview skills.

The key manipulation was the interviewers' supposed sex-role attitudes. Half the women were led to believe that the interviewer held traditional sex-role attitudes and was most interested in hiring a woman for easier jobs, and that she should be passive, gentle, unassertive, etc. The other half of the women were led to believe that the interviewer held nontraditional sex-role attitudes and was most interested in hiring a woman with equal work responsibilities, who was independent and assertive. The interviewers, of course, were confederates of the experimenter. They were not aware of what attitudes the women believed they held, and they were trained to act in an identical, neutral manner with all interviewees.

Did the women try to conform to the interviewer's supposed sex-role attitudes? Indeed, they did. First, in the traditional condition, the women arrived wearing more make-up and accessories (such as earrings), and independent raters judged them to be more attractive. Second, they talked less and were less likely to look directly at the interviewer while talking (both behaviors are typical of people acting in a more submissive manner). Third, they gave more traditional responses to an interview question asking them their orientation toward marriage and family. Thus, sex-role stereotypes had become self-fulfilling—the women's behavior conformed to their beliefs about the sex-role attitudes of the interviewers.

Self-fulfilling sex stereotypes in the classroom. Although considerably less well-known (and less cited) in the social psychological literature than the experimental studies I just reviewed, two classroom studies conducted around this time also demonstrated potentially self-fulfilling effects of sex stereotypes. The first (Palardy, 1969) examined whether teachers' beliefs

about sex differences in ability to learn how to read might be self-fulfilling. Palardy (1969) started with a pool of 42 first grade teachers and identified two groups: (1) One group believed that boys and girls learn to read equally well ("Group A") and (2) a second group believed that girls learn to read more quickly than boys ("Group B"). A self-fulfilling prophecy perspective led Palardy to predict that girls would outperform boys, but only when teachers held higher expectations for girls. Specifically, there should be little difference between boys' and girls' reading achievement in Group A (where teachers believed there was no difference), but girls should outperform boys in Group B (where teachers thought girls learned to read more quickly).

Five teachers from each group were then matched on demographics, teaching experience, and teaching methods. Reading readiness scores at the beginning of first grade and reading achievement scores at the end of first grade were obtained for 53 boys and 54 girls in Group A and for 58 boys and 51 girls in Group B. The reading readiness scores were nearly identical for all four groups. Thus, there were no real differences between boys and girls in either group, at least not initially.

The main outcome was end-of-year first grade reading achievement scores. (For the statistically inclined, these were submitted to a two [student gender] by two [teacher expectancy group: A,B] analysis of covariance [with IQ scores as the covariate].) Those scores resoundingly confirmed both of Palardy's predictions. By the end of first grade, there was no difference in reading achievement between boys and girls in Group A, but girls had higher reading achievement than boys in Group B.

A second study (Doyle, Hancock, and Kiefer, 1972) of 11 teachers and 245 students focused on three predictions: (1) On average, first grade teachers have higher expectations for girls than boys, (2) these different expectations are erroneous, and (3) erroneous expectations will be self-fulfilling. All three predictions were supported. Although there were no objective differences in boys' and girls' IQ scores, teachers estimated that boys had IQ scores averaging 99.9 and girls had scores averaging 104.5. The teachers had *underestimated* the IQs of nearly 59% of the boys and they had *overestimated* the IQs of nearly 57% of the girls. This pattern confirmed the first two predictions (erroneously higher expectations for girls).

Were these erroneous expectations self-fulfilling? Indeed, they were. Doyle et al. (1972) divided the students into two groups: those whose IQ scores were overestimated and those whose scores were underestimated. Despite slightly lower IQ scores, girls had higher reading achievement scores. In addition, the effect for discrepancy was highly significant: Students with a mean IQ of 98 (those in the overestimated group) actually outperformed those with a mean IQ of 107 (those in the underestimated group).

Social class stereotypes. Perhaps the most dramatic and well-known study of social class-based self-fulfilling prophecies was performed by Rist (1970). Rist observed a kindergarten class and found that by the eighth day of school, the teacher had divided her class into three groups—a supposedly smart, average, and dumb group. Each group sat at its own table (Tables 1, 2, and 3, respectively). However, the main difference between the students was not intelligence—it was social class. In comparison to the other students, the students at Table 1 came from homes that had greater incomes, were less likely to be supported by welfare, and were more likely to have both parents present, and the children themselves were cleaner and more likely to dress appropriately. There were comparable differences between the students at Tables 2 and 3. Table 1 was positioned closest to the teacher, and she proceeded to direct

nearly all of her time and attention to those students. In addition, Rist (1970) observed her to be friendlier and warmer to the students at Table 1.

Rist (1970) followed the class of kindergarten students through second grade. As did the kindergarten teacher, the first grade teacher assigned students to three tables (apparently according to her beliefs about the smart, average, and dumb students). All of the Table 1 (“smart”) students in kindergarten were assigned to Table A in first grade. However, students at Tables 2 and 3 in kindergarten were all assigned to Table B. Table C was reserved for the students the teacher believed were especially slow. In the second grade class, the students who had been assigned to Table A in first grade were all assigned to their own table (they were referred to as “Tigers”). Students who had been assigned to Tables B and C in first grade were assigned to a second table in the second grade class (referred to as “Cardinals”). None of the students from the first grade class Rist observed were assigned to the “slow” table (called “Clowns”).

Rist (1970) observed several important patterns: (1) The kindergarten teacher assigned students to tables on the basis of social class, but believed (or at least claimed) she did so on the basis of competence; (2) the teachers directed more of their time and attention to the children they believed were smarter; and (3) once labeled smart or dumb—which happened by the eighth day of kindergarten—the effects of that label lasted for years. Thus, Rist (1970) concluded that social class–based teacher expectations help create a “caste-like” system that benefits middle class children and undermines children from lower social class backgrounds.

The physical attractiveness stereotype. People associate all sorts of good things with being physically attractive, including intelligence, happiness, and kindness (Eagly, Makhijani, Ashmore, & Longo, 1991). However, they most strongly associate warmth, friendliness, and social skill with attractiveness (Eagly et al., 1991). A classic study in social psychology (Snyder, Tanke, & Berscheid, 1977) started with the assumption that this is an inaccurate stereotype, but suggested that the stereotype could become “true” through self-fulfilling prophecies.

To examine this possibility, Snyder et al. had to create a situation that had several characteristics. First, there had to be no initial differences between attractive and unattractive targets. Second, they needed to activate perceivers’ attractiveness stereotypes. Third, perceivers and targets would need to interact to provide an opportunity for those erroneous attractiveness stereotypes to become self-fulfilling. This was a tall order, and the elegance with which they did so attests to the creativity of this research.

The stickiest problem was how to activate the stereotype and still ensure that there really were no differences between attractive and unattractive targets. They accomplished this as follows. Male perceivers received a photograph of a female with whom they would soon be interacting. The photograph showed a woman who was either physically attractive or unattractive (this was determined through a pilot test, in which judges rated the women’s photographs). Presumably, the photograph activated the stereotype. In fact, however, male–female interaction partners were randomly assigned to the attractiveness condition. Thus, the photograph was *not* the same person that they were actually interacting with. Random assignment to “attractiveness” meant that there was little chance of there actually being important differences between supposedly attractive and supposedly unattractive women.

The male–female pairs, who were in different rooms, then had a telephone conversation. This allowed them to interact, but prevented the males from seeing their partners (of course,

this was necessary in order to maintain the males' belief that they were interacting with the person shown in the photograph).

Snyder et al. (1977) proceeded to map the interpersonal processes underlying self-fulfilling prophecies. First, did the males treat the supposedly attractive and unattractive women differently? They sure did. Both independent judges and the women themselves rated the male perceivers as warmer and friendlier when men believed their partner was attractive.

Second, did this create a self-fulfilling prophecy? That is, did the women believed to be attractive actually become more pleasant? Indeed, they did. Independent judges (who did not know anything about the attractiveness conditions) rated the women who were in the attractive condition as warmer and friendlier than the women in the unattractive condition. Thus, the women who the men believed were more physically attractive actually became more socially pleasant—thereby confirming the physical attractiveness stereotype. As Snyder et al. (1977) pointed out in their discussion, these results raised the possibility that all sorts of erroneous stereotypes might become “true,” not because the stereotypes were actually valid, but because people's erroneous beliefs tended to be self-fulfilling.

Self-Fulfilling Prophecies Are . . . Everywhere!

Self-fulfilling prophecies seemed to appear almost everywhere social psychologists looked for them. First, there was the early research on experimenter effects—self-fulfilling prophecies in, of all places, the scientific laboratory! Second, there was the classic, early research on teacher expectations—self-fulfilling prophecies in the classroom. Third, there was the early research on stereotypes—which documented that race, sex, social class, and physical attractiveness stereotypes all could be self-fulfilling.

Self-fulfilling competition. If all that was not enough, the research of this early period also showed that self-fulfilling prophecies appeared in all sorts of places. Kelley and Stahelski (1970) found it when people played the prisoner's dilemma game. This is a “game” in the sense that it is a highly structured situation, with clear rules and winners and losers—but it is not a “game” in the same sense as Monopoly or baseball. The prisoner's dilemma has been used for decades to study conditions under which people cooperate versus compete with other people. Although there are many variations, both players usually gain points when both cooperate (metaphorically, by working together it helps them both); both players usually either lose points or gain nothing when both compete (metaphorically, fighting with each other is bad); and if one competes while the other cooperates, the competitor usually gains a lot of points, whereas the cooperator loses a lot of points (metaphorically, the competitor exploits the cooperator).

Kelley and Stahelski (1970) first identified people predisposed to cooperate or compete with other people and then had them play this game, in all combinations (cooperator–cooperator, cooperator–competitor, competitor–competitor). The main question was, how would these different combinations of people play the game? For two combinations, the answer is pretty obvious: cooperator–cooperator pairs made cooperative moves nearly all of the time and racked up tons of points; competitor–competitor pairs made nearly all competitive moves and lost tons of points.

But what happened with the cooperator–competitor pairs? Although cooperators may have started out cooperating, they quickly realized that they were being exploited, because they kept getting victimized by their partner’s competitive moves. Thus, the cooperators often changed and started making many more competitive moves.

There also was a second question: Would either type of person become aware of the existence of two types of players (cooperators and competitors)? Cooperators did—they realized that some people made mostly cooperative moves and others made mostly competitive moves. Competitive players, however, never had the opportunity to discover this—because their competitive moves almost always evoked competition from their partners. Thus, all people looked competitive to the competitors.

In short, then, the competitors’ behavior is likely to lead them to adopt the view that the world is a hard, dog-eat-dog type of place. This belief, however, is self-fulfilling. Even when their partners were not out to get them, the competitors’ own behavior evoked selfish, competitive behavior from their partners. The potential relevance of this type of vicious self-fulfilling cycle to everything from unions negotiating with companies over wages and benefits to countries preparing for war or peace seemed obvious. If people (companies, unions, countries, etc.) believe “we have to get them before they get us,” they will not be predisposed to cooperate. Instead, they will be predisposed toward hostility and aggression. Once hostilities (strikes, political or economic sanctions, physical aggression) begin, even if the other side was prone to cooperate, they will usually feel a need to retaliate.

Self-fulfilling beliefs about others’ hostility. Are beliefs about others’ hostility self-fulfilling in contexts other than the prisoner’s dilemma or related situations involving negotiations? This, in essence, was the question examined by Snyder and Swann (1978a). First, they led perceivers to believe they were interacting with targets who either were or were not hostile. Of course, targets were randomly assigned to the hostility label condition, so that there probably were no actual differences in hostility between targets labeled hostile and nonhostile.

Snyder and Swann (1978a) then had them play a game, the object of which was to respond as quickly as possible (by depressing a telegraph key) to a signal provided by the experimenter. The person who responded fastest would be the winner. But there was an added twist: The players could also use a “noise weapon,” the purpose of which was to disrupt the other player’s reaction. There were six levels of noise this weapon created, ranging from mild to offensive and irritating. There was only one weapon, so the players’ access to the weapon would alternate every three trials (the perceiver could use it for three trials, and then the target could use it for three trials). The level at which this weapon was used constituted the main dependent variable.

Were perceivers’ beliefs about targets’ hostility self-fulfilling? Indeed, they were. First, perceivers who believed their opponent was hostile used considerably higher noise bursts than did perceivers who believed their opponent was nonhostile. Second, targets believed to be hostile reciprocated in kind, by giving their opponents higher noise bursts than did targets believed to be nonhostile. Thus, the targets actually acted in a more hostile manner.

There was, however, another aspect to this study that rendered this self-fulfilling prophecy even more dramatic. In addition to the hostile-label manipulation, Snyder and Swann (1978a)

also led perceivers to believe that their use of the noise weapon in the game either reflected their genuine predisposition toward being hostile (dispositional self-attribution, "I am using high levels of the noise weapon because I am an aggressive person") or reflected their circumstances (situational self-attribution, "I am using high levels of the noise weapon in response to my opponents' action"). Snyder and Swann (1978a) suggested that when a person makes a dispositional attribution for their hostility—that is, when they internalize their hostile behavior—such behavior is not likely to end when the game with this particular opponent ends. Instead, it is likely to continue in a subsequent interaction, even with a perceiver who does not consider the target to be hostile.

To test this idea, targets then played the game again, but this time with perceivers who were given no expectations about targets' hostility. Results clearly showed that targets who were believed to be hostile by the first set of perceivers continued to act in a more hostile manner (using higher noise bursts) with a second perceiver, but *only* if they had been led to internalize (make dispositional attributions for) their hostile behavior in the first round of the game. Neither the targets previously believed to be nonhostile nor the targets previously believed to be hostile but who were led to believe their hostility reflected circumstances (situational attributions) showed any evidence of hostility (all used fairly low levels of the noise burst).

These results were quite dramatic. They showed that interpersonal beliefs about hostility could be self-fulfilling. They showed that such effects were not restricted to the prisoner's dilemma game. And, perhaps most dramatically, they represented one of the clearest demonstrations of Merton's (1948) "reign of error"—the idea that an initially erroneous belief can, through self-fulfilling prophecies, take on a life of its own and become true. True, not just in the relatively transitory or superficial sense of a person acting in a particular way just with a particular perceiver who happens to hold a particular expectation, but in an enduring, potentially much more permanent way. The targets labeled as hostile who internalized their aggressive actions actually became more hostile people, even with perceivers who did not initially consider them to be hostile.

"Pygmalion Goes to Boot Camp." This was the title of an early article (Eden & Shani, 1982) investigating possible self-fulfilling prophecies among Israeli military trainees. The researchers used a sort of beefed-up Pygmalion-like expectancy induction procedure. They informed the military instructors that their trainees had been subjected to a series of highly validated tests that were extraordinarily successful at predicting trainees' future military performance. Furthermore, on the basis of these tests, they had assigned their trainees into one of three groups (high, regular, and unknown potential). There were three outcomes: trainees' performance, as measured by objective multiple choice tests and by an evaluation provided by a commander who was not the trainees' instructor and who was blind to their expectancy condition; a measure of trainees' attitudes toward the training course; and trainees' evaluations of the leadership qualities of their instructor.

Results were quite striking. The high-expectancy trainees scored much higher on performance tests than did either of the other groups of trainees. They also had more positive attitudes toward the training and evaluated their instructors much more positively. The effect sizes were quite large, indicating that these were no mere "statistically significant but practically trivial" patterns. The instructors' expectations had dramatically uplifted the high-expectancy trainees.⁴

(Preliminary) Conclusions

The classic and often dramatic early social psychology studies of self-fulfilling prophecies easily and naturally led to an infectious enthusiasm regarding expectancy effects in much scholarship about these phenomena. Self-fulfilling prophecies seemed to be everywhere psychologists turned their attention; they offered potentially important insights into social problems associated with stereotypes, prejudice, and inequality; and they offered equally potentially important insights into basic processes of social perception and social interaction. On top of all that, it was easy to tell an interesting and exciting scholarly research story on the basis of many of the classics, especially Rist (1970), Rosenthal and Jacobson (1968a, 1968b), Snyder et al. (1977), and Word et al. (1974).

As such, many researchers concluded that self-fulfilling prophecies were a powerful and pervasive social phenomenon. This infectious enthusiasm once pervaded much of the social psychological writing about expectancy effects, as can be seen from the following quotes:

(Referring to Rosenthal & Jacobson, 1968a, 1968b): "... the teachers' expectations had a dramatic impact on the actual performance of the spurters" (Gilbert, 1995, p. 131).

"... Events in the social world may be as much effects of individuals' beliefs as they are causes of these beliefs" (Snyder, 1984, p. 294).

"... Teachers' expectancies influence students' academic performance to a greater degree than students' performance influences teachers' expectancies" (Miller & Turnbull, 1986, p. 236).

"Constructivism asserts that we do not discover reality, we invent it" (Hare-Mustin & Maracek, 1988, p. 455).

"Once such an expectation is held about an individual, of course, self-fulfilling prophecy during social interaction should ensure that the hypothesis is behaviorally confirmed" (Skov & Sherman, 1986, p. 116).

"Attempts to understand the personal characteristics of others ... are complicated by the fact that one tends to find what one expects. This happens not only because information processing is selective, but also because expectancies cause one to act in ways that elicit behavior interpretable as confirming those expectancies even when the expectancies might have been mistaken" (Jones, 1986, p. 41).

"Self-fulfilling prophecies occur ... across a wide variety of situations. Although there are some circumstances that counter their occurrence, on the whole, biases in both the perceiver's and target's interpretations of the meaning of behavior and social norms for reciprocating behavior would seem to favor their development" (Fiske & Taylor, 1991, pp. 549–550).

"Thus the perceiver's expectancy has exerted an influence that extends far beyond the original interaction and can significantly affect the life of the target person—perhaps for the better, but as many who do this research fear, often for the worse" (Darley & Fazio, 1980, p. 879).

"It [social perception] often has significant and nearly direct influence on the perceived target. It creates social reality ... the hallmark of the cognitive perspective in social psychology is the constructive nature of social cognition" (Markus & Zajonc, 1985, pp. 212–213).

“... hundreds of experimental and naturalistic studies have provided strong evidence for expectancy effects in multiple domains, including schools. . . .” (Weinstein et al., 2004, p. 512).

When the first blush of social psychological research on expectancies is viewed in this light, it is no wonder that self-fulfilling prophecies seemed to be a ubiquitous social phenomenon, which, when understood, provided deep insights into how people socially constructed their own social realities. Furthermore, this love affair with expectancies was about to become even stronger. Not only did expectancies create actual social reality through self-fulfilling prophecies, but, even in the absence of self-fulfilling prophecies, expectancies seemed to often create powerful biases and illusions in the mind of the perceiver. That story, however, is told by the next chapter.

Notes

1. Chapter 8 appears in both lists because it focuses on a bias (in social hypothesis testing and information-seeking) that has been shown to lead to a self-fulfilling prophecy.

2. In much social psychological research on interpersonal expectations, “perceiver” refers to the person holding an expectation about a “target,” who is the person about whom the perceiver holds an expectation. So, for example, in Rosenthal’s various studies, experimenters and teachers were perceivers, and rats and elementary school students were targets. In most social interaction, of course, each person can be both a perceiver and a target. Nonetheless, the distinction is important, because most of the empirical research focused on examining effects of expectations flowing in only one direction—from perceiver to target (e.g., teacher to student, interviewer to interviewee, etc.).

3. This logic was not made explicit in exactly this form in any single individual empirical or theoretical paper. This, therefore, represents my own interpretation and synthesis of how this body of research came to be viewed. I do think, however, that this interpretation and synthesis both captures and distills the spirit of many of the ideas that were prevalent during this early period and that have periodically reappeared in much of the later writing about these issues (see, e.g., Devine, 1995; Fiske & Taylor, 1991; Gilbert, 1995; Hamilton, Sherman, & Ruvolo, 1990; Jones, 1986, 1990; Snyder, 1984; Snyder et al., 1977; Word et al., 1974).

4. For the statistically inclined, expectancies accounted for 28% to 73% of the variance in the outcomes.

5 The Extraordinary Power of Expectancies to Bias Perception, Memory, and Information-Seeking

SELF-FULFILLING PROPHECIES are not the only type of expectancy effect. Consider the following two examples. First, I have had colleagues tell me, in a sort of perplexed tone, that a particular paper they had recently reviewed in manuscript form, which they had not liked all that much, looked much better in the final printed journal format. Second, Dale Carnegie, in *How to Win Friends and Influence People*, suggests that most college students quickly learn that if they ace the first test or paper, it is much easier to ace the class (this was written in the 1930s, long before classes with 300, 400, or more students became commonplace). What could possibly connect such seemingly disparate events? Expectations. Specifically, both examples may show that people's expectations bias and color their interpretations of others' behaviors, accomplishments, and attributes.

In the case of the surprisingly improved manuscript, the institutional stamp of approval that comes with publication, plus the obviously more polished and professional look of a published journal article, may create a (relatively superficial) context of competence and quality (i.e., raised expectation) that colored my colleagues' evaluations of the papers. Carnegie's college students, too, might be taking advantage of an expectancy effect. Acing the first test or paper (in a class small enough for the professor to know who you are) will usually lead the professor to think that you are smart, motivated, and competent. If such an expectation also colors the professor's interpretations of your future work (e.g., you may receive the benefit of the doubt for marginal work), receiving an A in the class becomes considerably easier.

Thus, even if my expectation for you does not influence you at all, it may still influence how I see you. "How I see you" ("person perception," in social psychological lingo) involves

many things, including (but not necessarily restricted to) how I interpret, evaluate, judge, remember, and explain your behaviors, accomplishments, and characteristics. My expectations also may influence how I obtain information about you. And, in the era of the 1970s and early 1980s, social psychology's love affair with expectancies inspired many researchers to examine how interpersonal expectations influenced each of these aspects of person perception.

Next, therefore, I discuss some of the research from this era that most dramatically demonstrated these phenomena. As in the prior chapter, this review could not possibly be comprehensive because there are literally hundreds of studies that have addressed whether expectations bias perceptions. Instead, therefore, this chapter highlights a handful of the most well-known or influential studies from the early period of social psychology's love affair with expectancy effects (roughly 1970 to 1990). These studies have been selected precisely because they represent some of the best evidence from this time period regarding the ways in which expectancies bias perceptions. So, if there is any intentional bias in selecting studies to focus on, such bias should work primarily in the direction of overstating expectancy effects (the purpose of Chapters 4 and 5 is to convey some of the reasons for social psychology's early enthusiasm for expectancy effects—not to provide an even-handed evaluation of the research [which is coming in Chapters 6 through 9]).

Nonetheless, several types of studies are not included in this review. First, no studies addressing these issues after 1990 are discussed. Why 1990? The studies reviewed in this chapter, like those in Chapter 4, were selected in order to convey the enthusiasm for expectancy effects that often characterized the first blossoming of research that followed on the heels of Rosenthal and Jacobson's (1968a, 1968b) Pygmalion study. Later research will be discussed in later chapters. In addition, this review does not address any self-fulfilling prophecy studies—those studies are reviewed in Chapters 4 and 6 through 8. Last, studies are reviewed here only if their purpose was to study expectancy effects; studies whose main purpose was something else, but which might be construed as relevant to expectancy effects, are not included here.¹ Similarly, none of the myriad biases (see, e.g., Kahneman, Slovic, & Tversky, 1982; Nisbett & Ross, 1980) other than those produced by expectations are discussed here.

Expectations Bias Person Perception

STEREOTYPES BIAS PERSON PERCEPTION

Once again, much of the research on expectancy effects focused on stereotypes, for both sociopolitical and theoretical reasons. On the more sociopolitical side, at the time, stereotypes were widely viewed as a nearly unmitigated evil—they were supposedly irrational, rigid, inaccurate, resistant to change, etc. (Chapters 15 through 19 vigorously contest this view; nonetheless, this view was widespread). Stereotypes were believed to rationalize discrimination against all sorts of groups, such as women, African Americans, Latinos, gay men and lesbians, the physically handicapped, the psychologically disturbed, and many others. Many social psychologists wanted to *do* something about this unjustified state of affairs. And, being social scientists, a natural expression of such interest was to expose some of the

basic psychological processes through which stereotypes biased perception. On the more theoretical side, stereotypes constituted a natural source of expectations for individuals and, at least sometimes, a pretty powerful source of such expectations. Thus, stereotype-based expectations seemed to have considerable potential for biasing person perception in many ways.

Stereotypes as biased expectations: The classic view. The classic view—one found in both scholarly scientific writing and in popular culture (see Chapters 15 through 19 for a review)—is that stereotypes are incorrect, illogical in origin, based in prejudice, irrationally resistant to new information, exaggerations of real differences, and ethnocentric. In addition, they supposedly lead people to ignore individual differences, bias perceptions regarding particular individuals, and lead to self-fulfilling prophecies. This description of stereotypes is, I suspect, reasonably familiar to anyone reading this book. As documented in Chapters 2 and 15, social psychological perspectives on stereotypes generally adopted this characterization of stereotypes as a starting point.²

Such a perspective essentially *defines* stereotypes as biased perceptions. Beliefs about groups and their individual members, if incorrect and irrational, express the biases of the perceiver—not the attributes of the perceived. “Irrationally resistant to new information” and “rigid,” too, are essentially expectancy-based biases. In essence, resistance to new information and rigidity mean that people cling to their beliefs about groups regardless of how wrong they are—and even when directly confronted with the evidence that they are wrong. That is, the stereotype-based expectation leads to such biased interpretation of disconfirming evidence that the stereotype is maintained. “Exaggeration of real differences” is also an expectancy-based bias. Presumably the stereotype leads people to blow minor differences all out of proportion—that is, the expectation biases perception of the reality.

Racial stereotypes. This is all well and good, but so far, I have not presented any evidence that such bias actually happens. Even the evidence presented in Chapter 2 does not provide much evidence of bias. Katz and Braly (1933) only showed that there was widespread agreement on the attributes of various groups. They *believed* that such agreement reflected bias more than accuracy, but they actually had no evidence that this was the case. LaPiere (1936) tried to present evidence of bias, but (1) the evidence of biased perception was restricted to a small number of anecdotal quotes (thereby rendering it impossible for me to reach any conclusion regarding the generality or pervasiveness of such bias found in his study), and (2) his interpretation of his study emphasized a very different psychological process—that stereotypes rationalize prejudice, not that they bias interpretations of people’s behavior and attributes.

So is there any evidence that stereotypes bias interpretations and evaluations? Indeed, there is. In one oft-cited study, Duncan (1976) showed people a tape of two students getting into an argument that ended in a shove. There were four types of pairings of perpetrators/victims: African American/African American, African American/White, White/African American, and White/White. All perpetrators and victims were actually experimental confederates. The main result was that, when the African American student shoved the White student, 75% of the viewers described the action as “violent.” However, when the White student shoved the African American student, only 17% described the action as violent.

Duncan's (1976) study was later replicated in an integrated middle school (Sagar & Schofield, 1980). Children were read stories and saw pictures of two boys. One either bumped the other in the hallway, asked for the other's food, poked the other with the eraser end of a pencil, or used the other's pencil without asking. Both African American and White children rated that action as more threatening when the perpetrator was African American than when the perpetrator was White.

Sex stereotypes. The early research showed that this type of effect was by no means limited to race stereotypes. Goldberg (1968) investigated the role of sex stereotypes in judgments. He gave female college students written articles to evaluate. All students received the identical articles, with one difference—half were attributed to a male author and half were attributed to a female author (Joan vs. John McKay). The articles supposedly written by John were rated more positively than were articles supposedly written by Joan. Thus, women's evaluations of the articles were biased by their beliefs about the author's supposed gender.

Sex stereotypes can also bias people's explanations for men's and women's successes and failures. Deaux and Emswiller (1974) had perceivers explain the success of men and women attempting to accurately identify objects from a camouflaged background. Some objects were stereotypically masculine (wrench, tire jack, etc.); others were stereotypically feminine (double-boiler, mop [remember that this study was conducted in the early 1970s!]). The study was rigged so that the identifiers all performed much better than average. The main question Deaux and Emswiller (1974) addressed was: How will the success of men and women on this masculine and feminine task be explained? There was no difference in the explanation of men's and women's performance on the feminine task. But on the masculine task, the success of men was attributed much more to ability than to luck; the success of women, however, was attributed only slightly more to ability than to luck. Deaux and Emswiller (1974) suggested that this type of result could help explain enduring discrimination against women—if women who perform as well as men are seen as lucky, rather than skilled, women's supposed lack of skill could be used to justify maintaining their lower status.

Social class stereotypes. Darley and Gross (1983) examined the potentially biasing effects of social class stereotypes. Princeton students were led to believe that a fourth grade girl came from either a middle class suburban background or an inner-city impoverished background. Some then estimated her ability in liberal arts, reading, and math. These students showed little or no tendency to favor the student from the middle class background.

Others viewed a videotape of her taking a math test. All of these students saw the exact same tape of the exact same girl answering the exact same questions. Nonetheless, in comparison to when they believed she came from a lower class background, they rated her ability, cognitive skills, and motivation more highly when they believed she came from a middle-class background. They even claimed that the girl answered more questions correctly when they believed she was from a middle class background. Darley and Gross (1983) concluded that people's expectations bias their judgments when people feel they have clear evidence relevant to those expectations, but not in the absence of such evidence.

Stereotypes also bias memory. Thus far, the studies I have reviewed have all examined whether stereotypes influence judgments regarding targets. The next two studies showed that stereotypes also can influence what people remember about targets.

Cohen (1981) demonstrated that people more easily remember behaviors and attributes that are consistent with a stereotype than those that are inconsistent with that stereotype.

Perceivers in her study viewed a videotape of a dinner conversation between a husband and wife (they were actually husband and wife, but they were also experimental confederates trained by Cohen). Half the time, this conversation led perceivers to believe the woman was a waitress; half the time, the conversation led perceivers to believe the woman was a librarian. The remainder of the conversation conveyed an equal mix of librarian-like and waitress-like attributes and behaviors (e.g., librarian: wears glasses, has artwork in home, received history book as gift; waitress: no glasses, no artwork, received a romantic novel as gift).

Perceivers were then given a series of choices regarding objective aspects of the woman in the videotape (e.g., wore glasses . . . did not wear glasses). Their task was to select the correct description. Perceivers consistently remembered 5% to 10% more behaviors or features that were consistent with the woman's supposed occupation than behaviors or features that were inconsistent with her supposed occupation. For example, they were more likely to accurately remember that the "librarian" wore glasses and liked classical music, whereas they were more likely to accurately remember that the "waitress" had a beer and no artwork in her house (even though the tape was identical, showing the woman wearing glasses, liking classical music, having a beer, and not having artwork in her apartment). This pattern occurred across two studies and regardless of whether the memory test occurred immediately after the videotape or up to 7 days later. Thus, it appeared that category-based processing of social information led people to selectively remember stereotype-consistent information better than they remembered stereotype-inconsistent information.

Another highly influential study from this period investigated the role of stereotypes not only in biasing memory but also in "creating" memories. Snyder and Uranowitz (1978) first had college students read a supposedly true life history of a woman identified as "Betty K." This life history was carefully crafted to include elements consistent with college students' beliefs about both the typical background of lesbian women and the typical background of heterosexual women. For example, Betty was described as never having a steady boyfriend in high school, but going on dates, and as having a steady boyfriend in college who was mainly a close friend.

One week later, these same college students returned to the lab and read new material about Betty. Half discovered that she was now happily married; half discovered that she was now happily living with her lesbian lover. The main question Snyder and Uranowitz (1978) addressed was whether this new knowledge of Betty's sexuality influenced the students' memories of her life history. It did. First, much like Cohen's (1981) study, students accurately remembered more stereotype-consistent aspects of her life than stereotype-inconsistent aspects. Second, they also seemed to reconstruct her history to be more consistent with the stereotype than it actually was. That is, their errors of memory were more likely to be stereotype-consistent errors than stereotype-inconsistent errors. For example, when they believed that Betty was a lesbian, they might inaccurately remember her as having had few dates in high school. Snyder (1984, p. 267) concluded:

... the students in these investigations had allowed their preconceptions about lesbians and heterosexuals to dictate the way they wrote and rewrote the life and times of Betty K. . . . [A]s long as erroneous beliefs and assumptions about sexuality make it easy to remember evidence that supports these beliefs and assumptions and difficult to bring to mind evidence that questions them, people will continue to cling tenaciously to these erroneous articles of faith.

The mental illness label. In one of the classics of this early period, Rosenhan (1973) tested one of the most audacious hypotheses in all of psychology: that the insane are indistinguishable from the sane. Most of us, including most psychiatrists, clinical psychologists, and the lay public, probably believe that objective aspects of mental illness (e.g., hallucinations, obsessions, inappropriate emotional expressions, etc.) lead psychological and psychiatric experts to the diagnosis of insanity (pathology, psychosis, etc.). Rosenhan (1973) suggested that the causal process was *exactly opposite*—he proposed that the diagnosis comes first, and then the psychological and medical community's *interpretation* of the patient's behavior is entirely determined by the diagnosis. According to Rosenhan (1973, p. 251): "Psychiatric diagnoses, in this view, are in the minds of the observers and are not valid summaries of characteristics displayed by the observed." What does this have to do with stereotypes biasing person perception? It is essentially the same phenomenon—in both cases, the label (whether demographic or psychiatric) influences perceivers' interpretations and judgments of targets.

How did Rosenhan test this audacious hypothesis? He had eight sane people (i.e., people with no prior history of mental illness) admitted to psychiatric hospitals in order to see if the professional staff could identify them as sane. Why would they be admitted at all? To get admitted, all eight complained that they had been hearing voices. Upon admission, although they gave false identifying information, they then ceased displaying all intentionally false expressions of disturbed behavior and they did not intentionally alter any other aspect of their life history. Childhood experiences, family relationships, work experiences, etc., were all described as accurately as possible. Emotional experiences, both good and bad, were described as they had occurred.

If social reality typically has a large influence on social perception, the pseudopatients should have been readily detected. Because there was nothing pathological in the background of any of the eight people, Rosenhan (1973) claimed that this should have made it relatively easy for the doctors and nurses at these hospitals to figure out that the patients were fakers or, at least, sane. In contrast, however, because Rosenhan (1973) suspected that diagnostic labels evoke expectations that pervasively influence social perception, he predicted an exact opposite causal sequence. Rather than social reality causing social belief, social belief would entirely color people's perceptions of social reality. In this event, entirely normal behavior would be interpreted as evidence of insanity.

So what happened? All pseudopatients were kept from 7 to 52 days, with a mean length of stay of 19 days. When they were released, were any released because they were identified as sane? Not a one. All were released with a diagnosis of schizophrenia "in remission." Rosenhan (1973) argued at some length that this was no mere formality. No patient was identified as a faker; nor was there any evidence from hospital records to indicate that the staff considered any of the pseudopatients to actually be sane. Rosenhan (1973) argued that, from the staff's standpoint, "of course" an insane person who was functioning well enough to be released must be "in remission."

Even more striking, however, was the extent to which past and current events in the patients' lives were interpreted in the context of their psychiatric label. Rosenhan (1973, p. 253) describes this example:

A clear example of such translation is found in the case of a pseudopatient who had had a close relationship with his mother but was rather remote from his father during early

childhood. During adolescence and beyond, however, his father became a close friend, while his relationship with his mother cooled. His present relationship with his wife was characteristically close and warm. Apart from occasional angry exchanges, friction was minimal. The children had rarely been spanked. Surely there is nothing especially pathological about such a history. . . . Observe, however, how such a history was translated in the psychopathological context, this from the case summary prepared after the patient was discharged.

“This white 39 year-old male . . . manifests a long history of considerable ambivalence in close relationships, which begins in early childhood. A warm relationship with his mother cools during his adolescence. A distant relationship to his father is described as becoming very intense. Affective stability is absent. His attempts to control emotionality with his wife and children are punctuated by angry outbursts and, in the case of the children, spankings. And while he says that he has several good friends, one senses considerable ambivalence embedded in those relationships also. . . .”

As Rosenhan (1973) pointed out, it seems that this patient’s history was unintentionally distorted by the staff. There is nothing particularly ambivalent or unstable about his relationships. It seems that the diagnosis determined the interpretation of the history, rather than an evaluation of the history influencing the diagnosis.

All pseudopatients also took detailed notes of their experience. At first, they were secretive about doing so, because they feared this might give them away. However, they discovered that their fears were unfounded. Why? Any behavior, especially unusual behavior such as writing, was seen as a manifestation of their illness. The nursing records for one patient included the following: “Patient engages in writing behavior.”

Pseudopatients would sometimes pace the halls, out of boredom. “Nervous, Mr. X?” asked a nurse. The notes of the pseudopatients also include instances where, for example, the patients were mistreated by the staff in some way, resulting in an argument or altercation. Nurses and staff members never inquired about the source of such squabbles and apparently never entertained the possibility that something about the institution (its staff, its policies, etc.) might have had anything to do with the altercation. Instead, they always seemed to automatically attribute responsibility to the patient’s pathology. Thus, Rosenhan (1973) concluded that the diagnosis pervasively colored the institutional staff members’ interpretations of the pseudopatients’ behavior and life histories.

It’s Not Just Stereotypes!

OK, by now you are probably convinced that stereotypes can have profound and powerful influences on how people interpret, evaluate, remember, and explain the actions and attributes of those with whom they come in contact. We all belong to lots of social categories (male/female, ethnic groups, occupational groups, etc.), so it would seem apparent that stereotype-based expectancy effects profoundly influence how we see others and how they see us almost all the time.

That would seem to be pretty powerful by itself. But the effects of expectancies are by no means limited to stereotypes. Abundant evidence also showed that all sorts of other types of expectations could influence social perception.

Kulik's (1983) introversion/extroversion study. Kulik (1983) showed perceivers a videotape of an interaction between two people supposedly in a "getting acquainted session" as part of a study (they were actually experimental confederates). Perceivers were led to believe the target person was either introverted (restrained, enjoyed jazz, planned to pursue law, etc.) or extroverted (loud, smiled a lot, planned to pursue drama, etc.). During the videotape perceivers saw the target act in either an introverted (e.g., refraining from initiating a conversation with a stranger or new acquaintance) or extroverted manner (initiating a conversation with a stranger or new acquaintance). Thus, about one-quarter of the perceivers viewed the supposedly introverted target acting in an introverted manner, one-quarter viewed the supposedly introverted target acting in an extroverted manner, one-quarter viewed the supposedly extroverted target acting in an introverted manner, and one-quarter viewed the supposedly extroverted target acting in an extroverted manner.

There were two especially neat things about this study. First, the introverted and extroverted targets were actually the same person. Thus, any differences in judgments could not possibly result from actual differences in the personality of the target. Second, this allowed Kulik (1983) to use only two tapes of target behavior: one introverted and one extroverted. Thus, people in the introverted expectation, introverted behavior condition saw *the exact same* tape as people in the extroverted expectation, introverted behavior condition. Similarly, people in the introverted expectation, extroverted behavior condition saw *the exact same* tape as people in the extroverted expectation, extroverted behavior condition. This is important, because it means that, overall, the behavior of the targets in introverted and extroverted expectation conditions were identical. Thus, any differences in judgments between the introverted and extroverted expectation conditions could only have resulted from differences in the expectation, not from actual differences in behavior.

After viewing the behavior tape, perceivers then evaluated and explained the targets' behavior. Did the introversion/extroversion expectation influence how perceivers viewed the targets? Absolutely. First, the targets were seen as more extroverted in the extroverted expectation condition than in the introverted expectation condition, *even though they engaged in identical behaviors*. Thus, the expectation biased perceivers' interpretations of targets' behavior. Second, the expectation also influenced perceivers' *explanations* for target behavior. When the target acted in a manner *consistent* with perceivers' expectations (introverted expectation + introverted behavior, or extroverted expectation + extroverted behavior), perceivers explained the behavior as caused by the target's personality. However, when the target acted in a manner *inconsistent* with perceivers' expectations (introverted expectation + extroverted behavior, or extroverted expectations + introverted behavior), perceivers explained the behavior as caused by the situation. Thus, perceivers believed that introverts' personalities caused introverted behavior, but that extroverts' situation caused their introverted behavior. Similarly, perceivers believed that extroverts' personalities caused extroverted behavior, but that introverts' situation caused their extroverted behavior.

Rothbart, Evans, and Fulero's (1979) friendly/intelligent memory study. Much like the stereotype memory studies, this research showed that expectancies also enhance memory for expectancy-confirming information. One group of perceivers was led to believe that a group of targets was particularly friendly; another group of perceivers was led to believe that this same group of targets was particularly intelligent. Behaviors supposedly engaged in by these targets that reflected friendliness, intelligence, unfriendliness, or lack of intelligence

were then projected onto a screen (as well as a few behaviors unrelated to friendliness or intelligence).

After all target behaviors had been presented, perceivers were first asked to estimate the frequency of the various types of behavior. Sure enough, perceivers given an “intelligent” expectation estimated more intelligent than friendly behaviors; perceivers given a “friendly” expectation remembered more friendly than intelligent behaviors. Next, they were asked to remember as many of the individual behaviors that fell into each category. Again, they remembered more expectancy-consistent than expectancy-inconsistent behaviors.

Rothbart et al. (1979, p. 354) believed that, although this was not a study of stereotypes, the results had profound implications for understanding stereotypes: “To the degree that people selectively remember confirming instances, stereotypes may maintain themselves despite a relatively small proportion of confirming examples.” In other words, even if women are no more passive than men, even if African Americans are no more hostile than Whites, and even if Jews are no cheaper than Christians, many people will believe these groups have these attributes, because of their selective memory for confirming information.

Williams’ (1976) teacher expectation study. By the late 1970s, it was clear that teacher expectations can change students’ achievement through self-fulfilling prophecies (e.g., Rosenthal & Rubin, 1978). Do teacher expectations also bias teachers’ evaluations of students’ achievement? Specifically, do teachers evaluate high-expectancy students more positively than low-expectancy students, even when their objective achievements are similar? The results of Williams’ (1976) research suggested that they do.

Williams (1976) examined relations between teacher expectations and student achievement among over 10,000 high school students in Ontario and found clear evidence of perceptual bias. After controlling for IQ, previous grades, motivation, and socioeconomic status (SES), Williams found that teacher expectations significantly predicted grades, but not standardized test scores. In other words, teachers’ expectations seemed to influence their evaluations of students’ performance (grades) rather than students’ actual learning (as indicated by the standardized tests).

Expectations Bias Person Perception: Conclusions

The research on interpersonal expectations conducted in the 1970s and early 1980s showed that, even when expectations did not change *objective* social reality through self-fulfilling prophecies, they often changed subjective *perceptions* of social reality. Apparently, stereotype-based expectations often led people to evaluate, judge, interpret, explain, and remember other people’s behavior and characteristics in a manner consistent with those expectations. The behavior of African Americans was seen as more aggressive than identical behavior by Whites; the test performance of poor children was seen as lower than identical test performance by middle class children; once targets were labeled as “schizophrenic,” even highly trained professionals could not help but interpret the actions and life histories of perfectly normal people as reflecting severe pathology. Similarly, not only were stereotype-consistent behaviors more easily remembered, but also perceivers apparently reconstructed targets’ life histories to render them more consistent with perceivers’ own social stereotypes. The extraordinary power of stereotypes to bias judgments was a common theme in the social psychological

scholarship through the early 1990s (only a small sampling of which was discussed here—see, e.g., reviews by Brewer, 1988; Fiske & Neuberg, 1990; Fiske & Taylor, 1991; Hamilton, Sherman, & Ruvolo, 1990; Jones, 1986, 1990), and remains one today (e.g., Cuddy, Fiske & Glick, 2007).

And, of course, such effects were by no means restricted to stereotypes. Personality-based, behavior-based, and achievement-based expectations all had similar biasing effects on evaluation, interpretation, and memory.

You might be thinking, “Jumping Jehoshaphat! Interpersonal expectations influence everything. They are everywhere. They create self-fulfilling prophecies. They bias, color, and taint perception, judgment, and memory. They help sustain malicious social stereotypes. How much more influence could they possibly have on any aspect of social perception or social interaction?” And, although you might have asked that last question rhetorically, there is, in fact, at least one more way in which expectations can influence and color social perception.

Expectations Bias Information Gathering

How do we go about finding things out about other people when we interact with them? Do we engage in an even-handed, objective, nearly scientific assessment of their attributes? Or do we systematically search for information that confirms our expectations? Do we (even if unintentionally) channel social interactions such that we practically ensure that others will confirm our expectations? According to the early research in this area, the answer seemed to be a resounding “yes!” to the latter two questions.

A Brief but Important Tangent: What Constitutes Unbiased Information Gathering?

To understand whether expectations bias information gathering, one first needs to understand what would constitute unbiased information gathering. In the sciences, like in daily life, information is not usually sought randomly—scientists usually have one or more hypotheses that they are testing. A formal scientific hypothesis is in many ways not all that different from an informal, lay expectation (e.g., Kelly, 1955; Nisbett & Ross, 1980). Both involve people believing that if they look in the right places in the right ways they will be able to find some particular thing (event, person, phenomenon, etc.).

Consider a scientist who believes that cloud seeding can increase rainfall. Such a scientist might try seeding clouds on a bunch of overcast days to see how often and how much rain falls. If the hypothesis is correct, rain should fall more often and in greater amounts on days when clouds are seeded than on days that clouds are not seeded.

Of course, the hypothesis could be wrong. Perhaps cloud seeding is completely ineffective, or worse, perhaps *less* rain falls after seeding clouds. Comparing the frequency and amount of rain when clouds are or are not seeded provides a clear and objective way to find out whether cloud seeding works. Thus, the hypothesis that cloud seeding increases rain is fairly tested.

Now consider a soccer coach who has only one more starting position open, and who believes that Dori is a better player than Sara. To adequately test this hypothesis, the coach might play Dori and Sara one half of each of five games to see who plays better. If the

hypothesis is correct (and if all other things were roughly equal), the team would likely perform better in the five halves that Dori played. Perhaps they outscored the opposition by a total of five goals in those halves, but were outscored by five goals when Sara plays.

Of course, the hypothesis might be incorrect. If tested in this matter, an incorrect hypothesis would become readily apparent. For example, if the reverse occurred—the team outscored the opposition by five goals when Sara played, but were outscored by five goals when Dori played—it would seem likely that Sara was actually the better player. Again, the initial hypothesis is readily disconfirmable and can easily be revised to accommodate the new information.

The key point here is that objective, scientifically valid information gathering involves testing hypotheses in a fair manner. One does not only seek information that fits one's hypothesis. One does not rig the game so that the hypothesis is likely to be confirmed. One seeks to test the hypothesis in such a manner that the hypothesis is falsifiable—if your hypotheses are wrong, the evidence should be gathered in such a manner as to clearly show it. Indeed, one of the most common tenets in the philosophy of science (Popper, 1959/1968) is that scientific theories or hypotheses must be falsifiable. Methodologically, data must be collected in such a manner to allow for possible disconfirmation of one's hypotheses.

Social Hypothesis Testing

Do people seek the truth or do expectations bias information-seeking? How well do people's actual hypothesis-testing actions correspond to this scientific ideal? An innovative and groundbreaking series of studies by Snyder and Swann (1978b) suggested that the answer to this question is "not very well." They examined how people go about seeking social information in the context of an interview. In all studies, half the perceivers were asked to test the hypothesis that a particular target was an extrovert and half the perceivers were asked to test the hypothesis that a particular target was an introvert.

Snyder and Swann then gave each perceiver a list of 26 questions titled, "Topic Areas Often Covered by Interviewers," from which they were to choose 12 to ask the target. These 26 questions were a mix of extroverted questions (e.g., "What would you do to liven up a party?"), introverted questions (e.g., "In what situations do you wish you could be more outgoing?"), and neutral questions (e.g., "What are your career goals?").

The main dependent variable was how many of each type of question perceivers chose to ask. If people seek information in an objective and fair way, Snyder and Swann reasoned, their expectation for the target should have little effect on the questions they ask. But if, as they suspected, expectations bias information-seeking, people with an extroverted expectation for the target should select more extroverted and fewer introverted questions to ask, whereas people with an introverted expectation for the target should select more introverted and fewer extroverted questions.

The results consistently confirmed the prediction that expectations bias information-seeking. Across four separate studies, people with an extroverted expectation consistently asked, on average, slightly over seven extroverted questions and slightly under three introverted questions. In contrast, people with an introverted expectation asked slightly about five and a half introverted questions and under five extroverted questions.

Snyder and Swann (1978b) then searched for limits to this hypothesis-confirming bias—but failed to find any! First, they gave perceivers strong reasons to believe the hypothesis they were to test was wrong. For example, if perceivers were asked to test the hypothesis that a target was extroverted, they were informed that the target was a member of a group composed of about three-quarters introverts. Did this affect perceivers' information-seeking? Not at all. Not only did they continue to ask more hypothesis-confirming questions than hypothesis-disconfirming questions, but also this pattern was just as strong as in the initial study.

Next, Snyder and Swann (1978b) offered a \$25 prize to the person who selected the most informative questions with respect to determining the introversion versus extroversion of the target. This incentive did not reduce the hypothesis-confirming bias in information-seeking at all. Yet again, people testing the extroverted hypothesis asked more extroverted questions and fewer introverted questions, and people testing the introverted hypothesis asked more introverted questions and fewer extroverted questions.

Why is biased information-seeking a problem? Self-fulfilling prophecies (again)! The most obvious reason that biased information-seeking is a problem is that it will lead the information seeker to faulty, incorrect, and biased conclusions about the target. If I primarily probe for information that confirms my belief that you are an extrovert (or, by extension, intelligent or competent or knowledgeable, etc.), then I am likely to come away with an overblown impression of your extroversion (intelligence, competence, knowledge, etc.). Indeed, Snyder and Swann (1978b) specifically argued that this type of process may account for the persistence of popular misconceptions about other people in general, and for the perpetuation of clearly inaccurate social stereotypes in particular.

In addition to sustaining erroneous perceiver beliefs, however, the early research also showed that hypothesis-confirming information-seeking leads to self-fulfilling prophecies. That is, regardless of what they are really like, targets behave in a manner that confirms the perceiver's hypothesis. Snyder and Swann (1978b) showed this, too. In one of their studies, perceivers first selected their questions and proceeded to interview targets. Those interviews were recorded. A set of judges then listened *only* to the targets' responses (thereby avoiding any effects resulting from knowing the perceivers' questions). Targets in the perceiver-extroverted hypothesis condition were rated as more extroverted, confident, poised, and energetic than were targets in the perceiver-introverted hypothesis condition. Thus, not only did perceivers seek information in an expectancy-biased manner, but also they evoked actual, objective behavior from the targets that confirmed their (perceivers') hypothesis.

"Well," you may be thinking, "it is not really that surprising that outside observers rate someone who answers questions like 'What would you do to liven up a party?' and 'In what situations are you most talkative?' as more extroverted. After all, they hear a person talking about how lively and talkative they are!" However, not only do these types of questions lead targets to behaviorally confirm the hypothesis, but also, apparently, they lead targets to change their self-conceptions.

Fazio, Effrein, and Falender (1981) repeated Snyder and Swann's (1978b) procedures almost exactly, with one new twist. After answering the questions, they had targets rate themselves on introversion and extroversion. The result? Targets who had been asked the extroverted questions described themselves as more extroverted than did targets who had been asked the introverted questions. Apparently, then, not only do perceivers seek information in expectancy-confirming ways, and not only do perceivers' information-seeking

methods constrain targets' behavior to be more likely to confirm perceivers' expectations, but also targets internalize perceivers' beliefs about them as part of their self-concept! This pattern, therefore, further testified to the profound and powerful ability of interpersonal expectations to create their own reality.

Conclusions Regarding the Biasing Power of Expectations

Self-fulfilling prophecies, stereotypes, memory biases—one expectancy effect after another, and just when you thought there could not possibly be any more expectancy effects, there were attributional biases, teacher expectations biasing grades, and information-gathering biases. The extent to which expectations influence, change, and color (or, in the case of stereotypes, taint) our interactions with and perceptions of other people seemed to be nothing short of stunning. They pervasively color first impressions. They influence how we see other people, what we remember about them, how we explain their behaviors, and how we go about trying to figure other people out (not to mention changing how other people see themselves and actually behave).

The social psychological enthusiasm for expectancy-induced biases was at least comparable to, and perhaps exceeded, that expressed for self-fulfilling prophecies. Consider the following:

“Social perception is a process dominated far more by what the judge brings to it than by what he takes in during it” (Gage & Cronbach, 1955, p. 420).

“... these inferences [resulting from trait inferences, attributions, and implicit personality theories] may lead to inaccurate expectancies regarding the future behavior of the target” (Darley & Fazio, 1980, p. 870).

“... our beliefs pervasively color and bias our response to subsequent information, evidence, or argumentation” (Lord, Lepper, & Preston, 1984, p. 1231).

“Our beliefs and expectations have a powerful effect upon how we notice and interpret events” (Myers, 1987, p. 122).

“In sum, selective perception plays an active role in many negotiations, causing negotiators to perceive only the information that is in accordance with their biases” (Rubin, Kim, & Peretz, 1990, p. 129).

“Once cued, schemas³ affect how quickly we perceive, what we notice, how we interpret what we notice, and what we perceive as similar and different” (Fiske & Taylor, 1991, p. 122).

“All of the expectancy-driven processes described above ... are biased in the direction of maintaining the preexisting belief system ...” (Hamilton et al., 1990, p. 39).

“A particularly pernicious example of self-fulfilling beliefs and expectations, and the one most studied by social psychologists, is that of stereotypes and other negative beliefs about particular groups of people. ... If it is widely believed that the members of some group disproportionately possess some virtue or vice ... one is likely (in the absence of specific legal or social sanctions) to ... deprive or privilege group members in terms of opportunities to ... succeed or fail in accord with the beliefs and expectations that dictated their life chances” (Ross, Lepper & Ward, 2010).

(See also Chapter 15 for an additional collection of quotes focusing specifically on stereotypes.)

Preliminary Social Psychological Conclusions: On the Power and Pervasive Effects of (Widely) Inaccurate Expectations

The discussion of self-fulfilling prophecies in Chapter 4 and of expectancy-induced biases in the present chapter both ended with a string of quotes testifying to the inaccuracy and power of social beliefs, many from the most influential and well-known social psychologists in the field, publishing in the most widely read and influential research outlets. It is extremely easy to find such quotes because they reflect a dominant theme within social psychology. Furthermore, whole books have been written regarding the erroneous and biased nature of human social perception (Dawes, 1988; Gilovich, 1991; Kahneman et al., 1982; Nisbett & Ross, 1980). If I have done my job well in writing Chapters 4 and 5, you now have some insight into why enthusiasm for these effects once pervaded not only social psychology but also much of the social sciences. And, for the most part, it still does (e.g., Jost & Kruglanski, 2002; Ross et al., 2010; Weinstein, Gregory & Strambler, 2004).

This heavy emphasis on error, bias, and self-fulfilling prophecy was not shared by *everyone* who wrote about these issues. Nonetheless, the dissenting voices were few and far between, and were often presented as direct challenges to the prevailing emphasis on error, bias, and self-fulfilling prophecy (e.g., Funder, 1987; Kenny & Albright, 1987; McArthur & Baron, 1983; McCauley, Stitt, & Segal, 1980).⁴ It is clear that I am not alone in viewing the research and theory in this period as overwhelmingly emphasizing the power of inaccurate expectations. Consider one more round of quotes, by folks commenting on the prevalent research conclusions from this era:

“The image of the perceiver that emerges [from prior research on expectancies] is one of an individual who takes his or her stereotypes for granted and indiscriminantly applies them to members of the class he or she has stereotyped without a consideration of the unjustness of such a proceeding” (Darley & Gross, 1983, p. 32).

“... accuracy of perception implies a reality to be perceived, and the current resurgence of phenomenological approaches to social psychology tends to deny any such reality” (M. Cook, 1979, p. ix).

“... the literature has stressed the power of expectancies to shape perceptions and interpretations in their own image” (Jones, 1986, p. 42).

“... we are left with the uncomfortable conclusion that the give-and-take of social interaction cannot disconfirm prior impressions of others. In this respect at least, reality becomes irrelevant, if not denied” (Bond 1987, pp. 39–40, emphasis added).

“... the current Zeitgeist emphasizes purported flaws in human judgment to the extent that it might be ‘news’ to assert that people can make global judgment of personality with any accuracy at all” (Funder, 1987, p. 83).

“It does seem, in fact, that several decades of experimental research in social psychology have been devoted to demonstrating the depths and patterns of inaccuracy in social perception. . . . This applies . . . to most empirical work in social cognition. The thrust of dozens of experiments on the self-fulfilling prophecy and expectancy-confirmation

processes, for example, is that erroneous impressions tend to be perpetuated rather than supplanted, because of the impressive extent to which people see what they want to see and act as others want them to act . . .” (Jost & Kruglanski, 2002, pp. 172–173).

Even the most evangelical proponents of the inaccuracy and power of expectations, however, never claimed anything so silly or absolutist as “all expectations are always inaccurate” or “all inaccurate expectations are always self-fulfilling.”¹ Furthermore, although social psychological perspectives on expectancies rarely addressed the accuracy issue until the 1990s (not counting discussions of *in*accuracy), nearly all perspectives acknowledged at least some limitations to expectancy effects. Such limitations, however, were often presented as rare or unusual events, needing to overcome the common or “default” processes of self-fulfilling prophecy or bias (e.g., Fiske & Neuberg, 1990; Miller & Turnbull, 1986; Snyder, 1984). Even discussions of limitations, therefore, testified to the typical, although not absolute, power of expectancies.

The influence of such reviews cannot be overestimated. To this day, they constitute a primary resource regarding expectancy effects upon which many researchers rely. In short, the extraordinary emphasis on the power of expectations to create social reality that characterized the early reviews has become part of the distilled wisdom of social psychology.

Were Those Conclusions Justified?

“Well,” you may be wondering, “just because they emphasized the power and pervasiveness of expectancy effects does not necessarily mean that they *overemphasized* such effects. Perhaps their perspectives were simply true to the data!” Especially given the enthusiasm with which the research on expectancies was usually described, an enthusiasm I tried to recapture in Chapters 4 and 5, such a question is clearly warranted. If I emphasize how cold and snowy it is in Alaska, or how hot and wet it is in the Amazon, I am simply accurately and fairly describing an existing state of affairs. Perhaps the same can be said for the early conclusions regarding expectancies.

Whether such conclusions are valid, however, requires not merely an evangelical promotion of the early research—it requires a thorough and careful critical evaluation of that research. So, although it is clear that nearly all social psychological perspectives on expectancies during this first blush of research emphasized their inaccuracy and the power and pervasiveness of their effects, what remains unclear is whether such emphases were justified. Evaluating that, however, is no small task—and it is a task that begins in the next chapter.

Notes

1. For example, in Ross, Amabile, and Steinmetz’s (1977) classic quiz show study, subjects were randomly assigned to either answer (contestant) or ask (host) trivia questions. Both contestants and observers rated the hosts as more knowledgeable than the contestants. Why? Presumably, because (1) contestants could not answer many of the questions, whereas no evidence of the hosts’

ignorance was obtained, and (2) merely asking tough questions may make the asker seem smart. This is an interesting finding, and one important to attribution theories in social psychology. Is it relevant to expectancy effects? An expectancy interpretation would go something like this: Hosts have higher status than contestants; people have higher expectations for higher status people; therefore, their expectations led them to see the hosts as smarter than the contestants. None of this, however, was measured or manipulated. It is possible (I would argue, even likely) that expectancies had little to do with this effect—it was entirely a case of “behavior swamping the field.” Both contestants and observers witnessed the contestants repeatedly demonstrating their ignorance, and this heavily influenced their judgments.

The point of this long footnote, however, is not a detailed exposition of the study by Ross et al. (1977). That study is just an *example* of a whole class of research—stuff that is plausibly construable as related to expectancies, but that is just too indirect or tangential to warrant inclusion in this chapter. See Chapter 1 for an additional example of research that has been interpreted as relevant to expectancy effects, but which did not actually assess anyone’s expectations.

2. There has long been a minority of dissenters to this view (see, e.g., Ashmore & Del Boca, 1981; McCauley et al., 1980).

3. The social cognitive “zoo” is filled with all sorts of “animals” including, but not restricted to, expectations, beliefs, schemas, scripts, prototypes, stereotypes, categories, constructs, theories, and hypotheses. Although there occasionally may be important differences between these cognitive animals, all have sufficient similarity to be considered variations on expectations (see Jussim, 1991, for a more detailed presentation of this view).

4. This was not the case in educational psychology. Perhaps because of the hail of criticism leveled at the original Pygmalion study by some educational psychologists (e.g., Elashoff & Snow, 1971; see Chapter 3), educational psychology as a field developed a tradition very early on of caution and balance in considering the existence and power of expectancy effects. In contrast to the emphasis on error, bias, and self-fulfilling prophecy in social psychology, the reviews appearing in the educational psychology literature emphasized the accuracy of teacher expectations and the limited and modest power of most expectancy effects (e.g., Brophy, 1983; Brophy & Good, 1974; Cooper & Good, 1983; West & Anderson, 1976; Wineburg, 1987). This, of course, raises many questions, such as, who was right, the social or educational psychologists? Or maybe that is the wrong question. Perhaps both were right. That is, perhaps expectancies in educational contexts were typically accurate and not very self-fulfilling but were more likely to be inaccurate and self-fulfilling in other contexts. Addressing these questions, however, is not the point of this chapter (it is the point of the next several chapters).

5. Actually, this is not quite true—re-read the Skov and Sherman (1986) quote near the end of Chapter 4. Their article, however, was on hypothesis testing, not on self-fulfilling prophecies. However, the fact that they interpreted existing perspectives at the time as meaning that self-fulfilling prophecies would “ensure” fulfillment of an erroneous expectation testifies to the fact that I am not alone in interpreting those perspectives as emphasizing the power and pervasiveness of expectancy effects!

This page intentionally left blank

3 The Less Than Awesome Power of Expectations to Create Reality and Distort Perceptions

This page intentionally left blank

6 The Less Than Extraordinary Power of Self-Fulfilling Prophecies

CONSIDERATIONS BASED ON COMMON SENSE,
DAILY LIFE, AND A CRITICAL EVALUATION OF
THE EARLY CLASSIC EXPERIMENTS

Introduction to a Critical Analysis of the Early Research on Expectancy Effects

The theoretical emphasis of the reviews of expectancy research were well within the spirit of reviews of social cognition more generally (e.g., Fiske & Taylor, 1991; Markus & Zajonc, 1985), which long emphasized the power of social beliefs to create social reality. I have referred to this emphasis as the “strong social constructivist” position within social psychology (Jussim, 1991) because it emphasizes the extent to which people’s beliefs, attitudes, expectations, schemas, etc., create (“construct”) both objective social reality and their own subjective perceptions of that social reality. Such perspectives have often emphasized the inaccurate and error-prone nature of social beliefs while simultaneously suggesting that the accuracy of social beliefs is too complicated or uninteresting a question to address.

The next four chapters constitute a sort of intellectual forced march through several aspects of interpersonal expectancy research, all of which points to the same conclusion: The early emphasis on the power of interpersonal expectancies was unjustified. It was not justified by the classic early studies that remain highly cited today; it was not justified by other, less well-known research on expectancy effects; and it was not justified by the subsequent research on the same topics.

This can be readily seen from Table 6–1, which presents the overall average effect size for both self-fulfilling prophecies and biases, as obtained in every relevant meta-analysis

I could find. Except for the 0.52 effect among military personnel, all¹ range from about 0 to about .3 and do not show powerful or pervasive expectancy effects. For the statistically disinclined, an effect of .2 means that expectations substantially affect about 10% of targets, which, of course, is the same conclusion as that they *do not* affect 90% of targets. Or, put differently, a .2 effect means that high teacher expectations (compared to neutral ones) would raise a student's SAT scores by about 20 points.

Especially in light of the strong conclusions emphasizing their power highlighted in Chapter 4, how can the effects be as modest as shown in Table 6–1? That is the story of the next several chapters (Table 4–2 highlights precisely which expectancy phenomena are discussed in which chapter[s]). I revisit many of the studies already discussed and review some of the immediate follow-up research, in the spirit of documenting the justifiability of a far more modest set of conclusions regarding the power and pervasiveness of interpersonal expectancy effects.

INTRODUCTION TO THE LIMITED NATURE OF SELF-FULFILLING PROPHECIES

This chapter has two major sections that take a closer and more critical look at self-fulfilling prophecies in two very different ways. In the first section, I present a series of common experiences in daily life in which self-fulfilling prophecies either do not occur, occur to only a modest extent, or occur only infrequently. This is important both to help build at least a *prima facie* case against ascribing any sort of inevitability or great power to self-fulfilling prophecies and to link my research-based perspective on the limited power of expectancy effects to frequent everyday events. The second section revisits six of the most highly cited and classic self-fulfilling prophecy studies in order to evaluate just what they do and do not say about the power of expectancy effects.

On the (Non-)Inevitability of Self-Fulfilling Prophecies: Some Examples From Everyday Experience

SPORTS

Sports are filled with examples of expected winners failing and expected losers winning: Unheralded Marat Safin defeated Pete Sampras in the 2000 U.S. Open final; the 1969 Mets, who were widely predicted to wind up in last place, won the World Series; the New York Giants trounced the heavily favored New England Patriots in the 2008 Super Bowl. Of course, when one of two opponents has a clearly superior history, they will usually be favored, and, indeed, they will often win. This probably has a lot more to do with accuracy than with self-fulfilling prophecy. I doubt that the Yankees of the 1920s, 1950s, or late 1990s, or the Chicago Bulls of the early 1990s, or the San Francisco 49ers of the 1980s were great teams because people thought they were. It seems much more likely that people thought those teams were great because they played so well.

This is not to deny the existence of some degree of self-fulfilling prophecy in sports (e.g., Babad, Inbar, & Rosenthal, 1982; Trouilloud, Sarrazin, Martinek, & Guillet, 2002).

A coach who showers time, attention, and playing time on some players may increase their motivation and skill. A player who is ignored or heavily criticized, and who receives restricted playing time, might indeed become a weaker player.

At least part of the home field/home court advantage in many sports may come from the emotional high of being the target of the loud and vocal support of thousands of roaring fans. To the extent that this support reflects hope and enthusiasm, rather than an inaccurate expectation for the home team's success, this may not be exactly a self-fulfilling prophecy effect, but it is close. Informally, however, my experience listening to sports talk radio shows has been that fans do tend to overestimate their team's chance of success. Prior to the 2000 World Series, for example, the results of my official and highly scientific analysis (I listened in sporadically) of the people calling WFAN (a radio station in New York devoted almost entirely to sports) was that nearly all Yankees fans predicted the Yanks would win, and nearly all Mets fans predicted that the Mets would win. Perhaps fans often expect too much from their favored teams. If this translates into rooting for the home team, and if such support enhances the teams' performance, then a self-fulfilling prophecy may indeed partially explain the home court or home field advantage in sports.

The home court/home field advantages range from about 10% (e.g., in baseball, the home team wins about 55% of its games) to about 30% (e.g., in basketball, the home team wins about 65% of its games). This means that 70% to 90% of the games are determined by factors other than the home court/field advantage.² Thus, even if the *entire* home court/field advantage resulted from self-fulfilling prophecy (which is highly unlikely—consider field/court differences, effects of travel fatigue, etc.), it is clear that the overwhelming number of games are determined far more by team differences in skill, motivation, and preparation than by fan-based self-fulfilling prophecies. Regardless, sports has enough examples of favored individuals or teams losing that it provides strong *prima facie* evidence against any “inevitability” to self-fulfilling prophecies.

THE STOCK MARKET

My favorite example, though, is not sports—it is the stock market. One of the favorite social science everyday examples of self-fulfilling prophecies is the stock market. A very well-known and prestigious anthropologist once told me that my research on limitations to self-fulfilling prophecies was off-base because “of course everyone knows” that the stock market operates primarily on self-fulfilling prophecies. Similar claims can also be found in at least one popular social psychology textbook (Myers, 1999).

The process is supposed to work something like this. Investor expectations allegedly drive stock prices. If investors expect a stock to increase, many will buy, which drives up the price, which fulfills the original expectation. If investors expect a stock to fall, they sell, driving down the price, which fulfills the original expectation. This is a self-fulfilling prophecy (prices would not go up or down if it were not for the originally erroneous investor expectations). The kickers are, however, that (1) expert predictions regarding market performance are one of the best *anti*-self-fulfilling prophecy examples I know of; (2) although such stock self-fulfilling prophecies do happen, they do not happen very often; and (3) even when they do, it only accounts for relatively short-term fluctuations in stock market prices, not long-term trends.

TABLE 6-1

Where's the Beef? Average Expectancy Effect Sizes Typically Range from Small to Moderate Meta-Analysis		Topic/Research Question	Number of Studies	Average Expectancy Effect
<i>Self-Fulfilling Prophecy:</i>				
Rosenthal and Rubin (1978)		Do interpersonal expectations create self-fulfilling prophecies?	330	.29 ^a
Raudenbush (1984)		Do teacher expectations have self-fulfilling effects on student IQ?	18	.06
McNatt (2000)		Do manager's expectations have self-fulfilling effects on employees' performance?	6	.23
McNatt (2000)		Do military officers' expectations have self-fulfilling effects on trainees?	11	.52
<i>Bias in Judgment, Memory, and Perception:</i>				
Swim, Borgida, Maruyama, & Myers (1989)		Do sex stereotypes bias evaluations of men's and women's work?	119	-.04 ^b
Stangor and McMillan (1992)		Do expectations bias memory?	65	.03

Mazella & Feingold (1994)	Does defendant social category affect mock jurors' verdicts?		
	<i>Defendants':</i>		
	Attractiveness	25	.10
	Race (African American or White)	29	.01
	Social class	4	.08
	Sex	21	.04 ^b
Kunda and Thagard (1996)	Do stereotypes bias judgments of targets in the absence of <i>any</i> individuating information?	7	.25
Kunda and Thagard (1996)	Do stereotypes bias judgments of targets in the presence of individuating information?	40	.19

Notes. Effect size is presented in terms of the correlation coefficient, r , between expectation and outcome. All meta-analyses presented here focused exclusively on experimental research. "Individuating information" refers to information about the personal characteristics, behaviors, or accomplishments of individual targets. The effect size shown in the last column for each meta-analysis represents the average effect size obtained in that study. Effect sizes often varied for subsets of experiments included in the meta-analysis. Only meta-analyses of outcomes, not of moderators or mediators, are displayed.

^a This excludes the results of 15 studies on animal learning included in Rosenthal and Rubin's (1978) meta-analysis. In this book, expectations for animals are not considered to be "interpersonal" expectations.

^b A negative coefficient indicates favoring men; a positive coefficient indicates favoring women.

Market predictions. At any point in time one can find a large number of people making diametrically opposed predictions about the future direction of the stock market. You can easily discover this for yourself. Just read the prognostications of so-called experts appearing in the *Wall Street Journal* or the business section of the *New York Times*, or watch them directly on any televised stock market programs. Stocks been rising for a while? Some will say it's a great time to get in on the party and buy; others will say that stocks are overpriced and it is time to sell. Stocks been falling a while? Some will say they are cheap and now is a great time to buy; others will say that you better get out now if you do not want to lose your pants as well as the shirt that you already lost. It is obvious that they cannot all be right. It is just as obvious that they all cannot be self-fulfilling.

But the stock market provides an even better counterexample. Hundreds of supposed stock market experts offer their expertise (for a price) to the lay public through newsletters. Such newsletters typically include general information about investing and usually make specific recommendations regarding stocks or mutual funds to buy or sell.

These newsletters are an excellent predictor of stock market future performance (e.g., *Investors Business Daily*, 1996). So doesn't this show a self-fulfilling prophecy? Not at all, because they are a *contrary indicator*—that is, the higher the proportion of newsletters that are bullish, the more likely the stock market is to decline; the higher the proportion that are bearish, the more likely the stock market is to increase (*Investor's Business Daily*, 1996).

(Although it is beyond the scope of this book, I suspect you might be wondering how this could be. Aggregate expert recommendations *follow* the market rather than predict it. After a long period of stock market increases, most experts are bullish; after a long decline in stock prices, most experts become bearish. When nearly all are bullish, we are probably near a market top—that is, stocks are more likely to go down than up. When most are bearish, we are probably near a market bottom—and stocks are more likely to start going up than to continue down much further.)

The NASDAQ run-up of the late 1990s. Over small (a day or two) to even moderate length (a year or two) periods, investor-based self-fulfilling prophecies may indeed drive stock prices, both of individual companies and of entire markets (Siegel & Bernstein, 1998; Wijnenga, 1990). A classic example of this was the tech stock run-up of the late 1990s.

Although the Internet had been around in one form or another since the 1960s, it became widely accessible only in the mid-1990s. By 1995 or 1996, it had become pretty clear that the Internet was going to dramatically change how people communicate with one another and how at least some businesses were run. It was the “new” new thing. On top of that, cell phone technology also began to become widely popular at this time. As a result, billions of investment dollars—both private investment into start-up companies and public investment into the stock market—poured into communications, Internet, and technology companies.

The logic of a dramatic rise in tech stocks was, in some sense, pretty reasonable:

Technology is revolutionizing communications, computers, and business.

Companies providing that technology offer the most promising growth prospects.

Therefore, if I invest in those companies, I am likely to make tons of money.

The single largest concentration of such companies' stock trades are on the NASDAQ exchange in New York. From the beginning of 1996 to the beginning of 1999, the value of the

NASDAQ stock index went from about 1,200 to about 2,500. A 100% return in 3 years is a very good return. But it was nothing compared to what happened next. This logic, plus the recent success of NASDAQ companies, led to a bona fide market craze for tech stocks. And the NASDAQ proceeded to go from about 2,500 at the beginning of 1999 to over 5,000 in March 2000. That is a 100% return in 15 months and over a 300% return in just over 4 years.

This run-up in tech stock prices was insane. I am not just saying this in hindsight. This was *foreseeably* insane. Although I believe the evidence that, in general, buying stocks and holding them is a good investment strategy, this run-up was so crazy that, in early 2000, I removed most of my retirement account from the more aggressive tech-heavy fund I had been investing in. Thank goodness for me—my retirement portfolio was flat in 2000, whereas many of my friends and colleagues saw declines of 30%, 50%, and sometimes more.

This is why it was foreseeably insane. Stocks represent shares of ownership in a company. Therefore, the value of the shares should vary with the value of the company. More valuable company, more valuable shares. Why should the stock price of a company that is not growing and has no future prospects of growing increase? There is no reason. Growing companies become more valuable, which increases the value of their stocks. To justify a 100% yearly return on a stock price, therefore, companies must, on average, grow at about 100% per year. Consistently. Year in and year out. A 70% or even 50% growth year might be OK, if it was surrounded by years of 130% or 150% growth.

Although a very small number of companies have managed to achieve this level of growth for a few years, no company has ever achieved this level of growth for an extended period of time. New companies—like all those new Internet start-up companies of the mid- and late 1990s—were more likely to go out of business than to grow profits at 100% per year. Even the highest of high-growth companies rarely can sustain a yearly growth rate of 15%—let alone 100%—for more than 10 years. (Just do the math; the faster a company grows, the bigger it gets, by definition. The bigger it gets, the more it has to grow, in absolute terms, to sustain the same *rate* of growth. Therefore, the bigger you are, the more it takes [and the harder it gets] to sustain a very high growth rate. For example, 100% growth for a company that made \$1 million last year means adding a second *million* dollars in sales. To double again, it needs to add \$2 million in sales. To sustain this for 10 years, it needs to add \$2 *billion* in sales in its 10th year. In absolute terms, this company has to grow literally *one thousand times faster* in year 10 than in year 2 in order to sustain a 100% growth rate. And so on.)

The yearly profit growth rate for most large publicly traded companies typically falls in the 5% to 10% range. The high-growth, well-established tech companies in the 1990s—like Microsoft, Cisco, and Oracle—grew sales and profits by about 30% to 50% per year. Almost no company can sustain 100% growth. And, amazingly, many of the Internet and other high-tech companies were not only *not* growing profits at 100%—they were *losing* money. It was April 2001 when I first drafted this chapter, and Amazon.com (one of the darlings of the Internet craze, whose stock price jumped from about \$5 in early 1998 to over \$110 in late 1999), for example, still had not made a dime's worth of profit (which may help explain why its stock price fell in 2001 to under \$15; note: it eventually developed a decent business model and became profitable).

Thus, it is probably reasonable to consider the NASDAQ run-up of the late 1990s the result, at least in large part, of a self-fulfilling prophecy. Investors believed that tech was the wave of the future, they poured money into all sorts of tech companies, and the stock prices

of those companies doubled and then doubled again, in a few short years. Classic self-fulfilling prophecy. If so, it was important, at least for some people. The few who bought high-tech stocks in the early or mid-1990s and sold them in 1999 or early 2000 made a lot—and I mean a lot—of money. I suspect that such people are few and far between, however, because it is *extremely* difficult to know when to buy and when to sell.

More people, I suspect, got hammered by the subsequent fall. Oh yes, NASDAQ prices eventually came back down to earth. When I first wrote this chapter the NASDAQ had been around 2000 for the last several months (it eventually fell below 1,200 and has been between 2,000 and 2,500 for most of the last several years). Many of the Internet start-up companies have gone out of business—leaving their shareholders' stocks worthless. Many of those companies that did not go out of business saw their stock prices drop by 80%, 90%, or even more. Because people tend to enter the market after a long upward run ("Hey, look how easy it is to make a 50% return in the stock market!"), I suspect that more bought than sold near the peak in 1999 and early 2000. These people, of course, lost a lot of money.

So, in the short to intermediate run (a few weeks, a few months, even a year or two), stock prices probably do indeed change as a result of the self-fulfilling expectations of investors. The key phrase here is "short to intermediate term." The overwhelming *long-term* (5 years, or more) influence on stock prices, however, has nothing to do with investor expectations. It has to do with company profit growth (Siegel & Bernstein, 1998). The more money companies make, the higher the value of their stocks. Indeed, over decades, major market indexes, such as the Dow Jones Industrial, the S&P 500 (which tracks the stock price of 500 of the largest U.S. companies and accounts for about 70% of all U.S. stock market trading), and the Wilshire 5000 (which tracks all U.S. stocks) march almost in lock-step with company growth. Self-fulfilling prophecies (and many other relatively short-term factors, such as war, political scandals, etc.) can indeed perturb this pattern over the short or intermediate term. But when stock prices rise dramatically faster than does corporate growth (e.g., the 1920s, 1960s, 1990s), stock prices are usually headed for a fall and/or long period of weak growth (e.g., the 1930s, 1970s, and 2000s; see Siegel & Bernstein, 1998). This is not an expectancy effect. Over the long term, economic reality, far more than investors' self-fulfilling expectations, determines stock prices.

BANK INSOLVENCY

Merton (1948) introduced his analysis of the self-fulfilling prophecy with a metaphor about the "Last National Bank"—a perfectly healthy bank that became insolvent in the 1930s because of a depression-inspired bank run. This is one of the most famous self-fulfilling prophecy examples in the entire literature, because it is so compelling.

In the 1980s, however, hundreds of savings and loan institutions throughout the southwest United States became insolvent. The situation so seriously threatened the economic health of the country that Congress dedicated billions of tax dollars to bail them out. But this was no self-fulfilling prophecy. The problem was not bank runs. Indeed, the bankers *expected* their (now obviously excessively risky and speculative) loans and investments to produce handsome profits. If expectations were typically powerful and pervasive, even if the bankers were initially wrong, their beliefs should have come true. They didn't. And we all had to pay (through our taxes) to help bail out this sort of anti-self-fulfilling prophecy.

(A sadly ironic and relevant epilogue along the same lines: I am putting the near-final touches to this chapter in 2008, just as a new economic crisis, involving subprime loans and frozen credit markets, is requiring another government bailout. The financial wizards on Wall Street thought they were making a killing. Instead, they [those at Bear-Stearns, Lehman Brothers, and the like] lost their shirts. So much for self-fulfilling financial expectations. . . .)

COMMON SENSE WRAP-UP

I could go on (people from humble beginnings who become successful; injured or ill athletes who are told they can never compete again and then not only compete but win; students who enter graduate school with high GRE scores, great undergraduate records, and sky-high faculty expectations who drop out; and so on). But I hope it is not necessary.

Clearly, we all have different personal experiences and will differ in how we interpret the experiences that we share. I fully understand that I might interpret something from daily experience as accuracy that you interpret as bias or self-fulfilling prophecy. But I also think that, with minimal effort, most of us can bring to mind enough examples of inaccurate expectations that were not fulfilling to at least raise considerable doubts about the viability of claims attributing great power or near inevitability to self-fulfilling prophecies. Of course, I am not suggesting that such doubts rely exclusively on an appeal to common sense or personal experience.

A Critical Evaluation of the Early Self-Fulfilling Prophecy Classics

In this section, I revisit six of the early classic studies of self-fulfilling prophecies. As in previous chapters, this review is selective rather than comprehensive. Indeed, it is *very* selective. With such a small sampling of studies, you would be justified in wondering whether I purposely selected studies that fit my own preferred conclusions regarding expectancy effects. In fact, however, I have done just the opposite.

These six classic studies constitute highly cited pillars of virtually any perspective that emphasizes the power and pervasiveness of expectancy effects. All are typically cited in major reviews of self-fulfilling prophecies, and one or more often pop up when, for example, researchers want to argue for the practical importance and influence of stereotypes and prejudice in leading to discrimination and inequality. So, the articles included for review in this chapter are only a small, select, and biased sampling of research on self-fulfilling prophecies—but that bias is entirely in the direction of including studies commonly interpreted as emphasizing the power and pervasiveness of self-fulfilling prophecies.

Why would I selectively focus on such studies? Because, as far as I can tell, *even such studies* fail to justify an emphasis on the power and pervasiveness of expectancy effects. Thus, another piece of the puzzle falls into place. Not only does daily life provide numerous examples of limited or nonexistent self-fulfilling prophecies, but also even the studies most frequently cited in testaments to the power of self-fulfilling prophecies actually provide little such evidence. Although this does not complete the puzzle (for that you will have to read the rest of this book!), it is one important piece. That said, let's revisit the classics!

ROSENTHAL AND JACOBSON (1968A, 1968B)

Chapter 3 reviewed this study at length, and amply demonstrated how, even if you take its results at face value (which may or may not be justified), those results themselves were quite modest. There was no assessment of teacher expectation accuracy, the overall self-fulfilling prophecy effect size was quite small, and statistically significant self-fulfilling prophecies did *not* occur in 8 of 11 grades studied.

RIST (1970)

Rist's (1970) observational study (detailed in Chapter 4) was at one time particularly influential perhaps because it seemed to fill in some of the scientific and political blanks left by Rosenthal and Jacobson's (1968a, 1968b) Pygmalion study. Specifically, it seemed to demonstrate that teachers greatly underestimated and mistreated (primarily by ignoring) students from lower social class backgrounds and that these inaccurate negative expectations were so powerfully self-fulfilling that they created an academic caste system.

Did Rist really find evidence of self-fulfilling prophecies? The differences Rist (1970) observed in teacher treatment of middle class versus poor students would be inappropriate and unjustified even if there were real differences in the intelligence of the children at the different tables. Nonetheless, despite Rist's (1970) conclusions, *the study provided no evidence of self-fulfilling prophecy*. None.

Although Rist provided a wealth of observations concerning teacher treatment, he provided few regarding student performance. Differential treatment alone is not evidence of self-fulfilling prophecies. Differences in student outcomes are also needed. The one student outcome measure that Rist (1970) provided was students' IQ scores. In contrast to the self-fulfilling prophecy hypothesis, there were no IQ differences between the students at the different tables at the end of the school year. Thus, although the teacher may have held very different expectations for middle- versus lower class students, and even though the teacher may have treated students from different backgrounds very differently, this did not affect students' IQ scores.

This does not mean self-fulfilling prophecies did not occur. Perhaps they did. But no other evidence regarding objective measures of students was reported. Therefore, it is fair to describe the study as failing to provide evidence of self-fulfilling prophecies (which could be because they did not occur or because they did occur but Rist failed to find them).

What about the caste system? As in kindergarten, in first grade the students were again placed at tables supposedly reflecting achievement. All of the students from kindergarten Table 1 (the high table) were placed at the first grade Table A (high group). Nearly all of the students from kindergarten Tables 2 (middle) and 3 (low) were placed at first grade Table B (middle group). One of the students from the kindergarten class was placed at the lowest first grade table, Table C. Although students from the high-ability table remained at the high-ability table, the students from the middle- and low-ability tables in kindergarten were combined into one middle-ability table in first grade. Thus, if seating assignment is the criterion for evaluating whether a social class "caste system" existed, at this first transition, overall differences among students based on reading table assignment had declined.

By second grade, the students from Table A were assigned to the "Tigers" (high group) and students from Tables B and C were assigned to the "Cardinals" (middle group). None of

the students from the first grade class were assigned to the “Clowns” (low group). In addition, that year, two students from the “Tigers” were moved down to the “Cardinals” and two students from the “Cardinals” were moved up to the “Tigers.” Although the groups created by the kindergarten teacher did remain somewhat intact from year to year, by the end of second grade, initial differences (as indicated by seating assignments) between students had decreased. Thus, although Rist (1970) interpreted his study as demonstrating that expectancies contribute to a caste-like system based on social class, his actual results show considerable fluidity between the supposed castes.

A colleague once described the Rist (1970) paper as “a real tear-jerker,” and I can’t help but agree. But Rist provided little evidence of self-fulfilling prophecies, and no evidence that teacher expectations contributed to a rigid caste system based on social class.

Replication. There have been no published attempts to replicate Rist’s (1970) study exactly—that is, no observational studies of self-fulfilling prophecies within a single class of students over a school year or more. However, quite a few studies have addressed the same and highly related issues (e.g., the role of student social class in teacher expectations) using more rigorous and quantitative procedures. This research has consistently shown that teachers perceive social class differences between students because there really are differences in their performance and achievement, not because teachers are unduly biased by students’ social class backgrounds. Student social class has little or no influence on teacher perceptions over and above objective measures of performance such as standardized test performances or grades (Jussim & Eccles, 1995; Jussim et al., 1996; Madon et al., 1998; Williams, 1976).

Williams’ (1976) study of over 10,000 high school students is typical of the manner in which the follow-ups (Jussim & Eccles, 1995; Jussim, Eccles & Madon, 1996; Madon et al., 1998) provided a much more rigorous analysis of the role of social class in teacher expectations than did Rist’s (1970) study of a single kindergarten class. Williams (1976) found that teachers held higher expectations for students from higher socioeconomic backgrounds. However, these differences in teacher expectations evaporated after controlling for students’ previous levels of performance. This means that, rather than student social class biasing teacher expectations, teachers accurately perceived genuine differences in achievement among students from differing socioeconomic backgrounds.

The less well-known Williams (1976) study (and our own research) is much stronger than Rist’s (1970) study on almost all important scientific grounds—Rist relied primarily on his own subjective and potentially biased observations, whereas Williams and my team relied on school records and questionnaires; Rist focused on 30 students, whereas Williams focused on over 10,000 students (we focused on about 1,000 to 2,000, depending on the study); Rist claimed to provide strong evidence of self-fulfilling prophecy but actually provided none, whereas Williams rigorously tested for self-fulfilling prophecies and failed to find any (we found small ones overall). Although social class may sometimes lead to self-fulfilling prophecies, with respect to drawing scientific conclusions based on evidence, the subsequent research deserves dramatically more weight than Rist’s (1970).

WORD, ZANNA, AND COOPER (1974)

This is the White/African American interviewer/interviewee study detailed in Chapter 4. Although a classic, this study’s limitations suggest caution might be warranted in considering its implications.

Limitations. For example, ethnic stereotypes were never measured. Perhaps the self-fulfilling prophecy was triggered, not by perceivers' stereotypes, but by their prejudice (disliking) of African Americans (see Park & Judd, 2005, for a similar perspective). Word et al. (1974) ran a pilot study that documented that other Princeton students were indeed prejudiced against African Americans. Alternatively, the source of the self-fulfilling prophecy may have been neither stereotypes nor prejudice. It may have been anxiety. People often feel anxious when interacting with members of a different ethnic group, especially when the groups have a long history of conflict (e.g., Stephan & Stephan, 1985). Clearly, the source of the White interviewers' different behavior toward White and African American interviewees remains to be pinned down.

Even more important, it is not clear that the findings from Word et al. (1974) readily generalize to other interracial interactions. The study was conducted at a virtually all-White, historically (at the time) all-male elite Ivy League university (Princeton). Therefore, whether the White students in this study acted in a similar manner and held similar stereotypes and prejudices as, for example, White students at more diverse universities, then or now, remains an open, unanswered empirical question.

Replication. I completed this chapter in 2008. This is worth mentioning, because in the 34 years since this study was published, no attempts to replicate it have appeared in the published research literature.³ Whether the study *could* be replicated is unclear.

There are reasons to suspect such replication would be difficult. First, although prejudice is alive and well, the United States may be a considerably more egalitarian society in 2008 than it was in 1974. Especially at colleges and universities, diversity and multiculturalism are often highly valued. Second, with respect to wealth, social status, and background, there are few colleges or universities around the country that draw as elite a group of students as does Princeton. Such a background, especially in 1974, may have been especially conducive to prejudice and, indeed, to the type of elitism likely to evoke a self-fulfilling prophecy.

This is, admittedly, conjecture. In the absence of replication, anyone can speculate as freely as they like about the generalizability of this study. If some researchers want to cite it as clear and convincing evidence of the self-fulfilling nature of racial stereotypes, they are free to do so—there is no contradictory evidence. If other researchers want to suggest that the generalizability of this study is likely to be highly limited and its results difficult to replicate, they are free to do so—there is no evidence to contradict this, either.

That is the problem with lack of replication. All studies have important limitations, even classics such as Word et al. (1974). In the absence of replication, either the results of the study or its limitations can freely be emphasized in social science writing and discourse about stereotypes and self-fulfilling prophecies.

My own take is to be conservative—to consider the study suggestive, but to emphasize its limitations pending replication. In justification for this cautiousness, I cite two non-self-fulfilling prophecy research findings. First, the Darley and Gross (1983) study that demonstrated that social class stereotypes biased evaluations of a fourth grade girl taking a math test (described in Chapter 5) was, like the Word et al. (1974) study, conducted at Princeton. Using much the same materials (they obtained them from Darley), Baron, Albright, and Malloy (1995) attempted to replicate the study at two New England universities and failed to find any bias (these studies are discussed in more detail in Chapter 9).

Similarly, when African Americans expect to interact with a prejudiced White person, they apparently work especially hard to create a pleasant interaction—and often succeed at doing so (Shelton, 2000). Thus, awareness of the potential for perceivers to be prejudiced may reduce the potential for self-fulfilling prophecy (see also Hilton & Darley, 1985). Neither Shelton's nor Baron et al.'s research was a self-fulfilling prophecy study. Nonetheless, at minimum, they should raise some doubts about the generalizability and ease of replicability of the classic self-fulfilling prophecy study by Word et al. (1974).

SNYDER, TANKE, AND BERSCHIED (1977)

This is the classic male–female attractiveness self-fulfilling prophecy study described in Chapter 4. It is one of the most well-known self-fulfilling prophecy studies in all of social psychology and was cited by every one of the reviews emphasizing the strong constructivist perspective discussed in Chapter 6.

Was an inaccurate stereotype the source of perceivers' erroneous expectations? So, did this study show powerful effects of people's erroneous stereotypes? Let's start with the beginning. Why did males develop erroneous expectations? It was not because they held demonstrably erroneous beliefs about differences among attractive and unattractive women. Snyder et al. (1977) did not assess the accuracy of such beliefs. Indeed, because they subscribed to the prevailing wisdom that stereotypes were generally inaccurate, they simply assumed that the physical attractiveness stereotype, too, was inaccurate.

Such a presumption, however, goes too far. People who are physically attractive are indeed more socially skilled than others who are less attractive (e.g., Goldman & Lewis, 1977; see meta-analyses by Eagly, Makhijani, Ashmore, & Longo, 1991; Feingold, 1992). Thus, expecting physically attractive people to be more pleasant than less attractive people would, on average, be more accurate than expecting no difference.

Why, then, did the male perceivers in this study develop erroneous expectations? Because they based their expectations *on a photograph of a person who was different than the person they interacted with*. By randomly assigning subjects to attractiveness conditions, Snyder et al. (1977) artificially rendered the relationship between perceived attractiveness (which was randomly assigned) and social skill to be zero. In this context, therefore, basing beliefs on attractiveness led to a false expectation—not because belief in a connection between attractiveness and pleasantness was false, but because the attractiveness of the person in the photo did not necessarily correspond to the attractiveness of the person with whom the males interacted. It seems, therefore, that the males in the Snyder et al. (1977) study developed erroneous expectations by doing something that in nearly all other initial interaction contexts would lead them to be as accurate as possible under the circumstances—that is, utilizing attractiveness information as a basis for expectations regarding social skill.

Self-fulfilling stereotype? But there is an even more serious potential weakness to this study. Specifically, it is not clear that there was any activated stereotype or “prophecy.” “If it wasn't because of their attractiveness stereotype,” I can almost hear you asking, “why, then, did the males act differently toward the supposedly more attractive women?” Because they were interested in impressing a pretty woman. And why might 19- and 20-year-old college men be interested in impressing a pretty woman? Do you really have to ask? Whether this plausible

alternative explanation for their findings is true is unknowable from their data. What is true, however, is that their procedures were not capable of ruling it out. Bottom line: Although it is clear that males were warmer to supposedly attractive women, it is not clear that stereotype-based expectations triggered males' warmth.

The irretrievable effect size. In addition, although this study has been widely cited as a testament to the power of beliefs, and especially stereotypic beliefs, to create social reality, Snyder et al. (1977) provided no information about the *size* of the self-fulfilling prophecy effect. The independent judges rated the women on 21 traits related to the attractiveness stereotype (sociable, poised, warm, etc.). Snyder et al. (1977) performed a multivariate analysis of variance (MANOVA), which determined whether all means were equal across attractiveness conditions. This analysis was statistically significant, which meant that at least 1 of the 21 means differed across groups.⁴ Typically, researchers perform additional statistical analyses after obtaining a significant MANOVA to figure out precisely *which* means differed by condition. Snyder et al. (1977) did not report this, which was particularly unfortunate from our modern standpoint, because effect sizes can be estimated on the basis of most analyses performed on one dependent variable at a time (*t* tests, univariate analyses of variance, correlations, etc.). Instead, they pointed out that the means on the 21 variables were in the predicted direction 17 of 21 times and performed an analysis indicating that that was not likely to happen by pure luck. This is fine as far as it goes, but we were still left without any clear information regarding the size of the effect.

Generalizability to lasting relationships? Last, the study focused on a short-term interaction between strangers. Thus, it provided no information about the extent to which self-fulfilling prophecies were likely to have enduring effects within the context of long-term relationships.

Replication. There has been only a single direct attempt to replicate the findings of this study. It failed. Andersen and Bem (1981) had androgenous or sex-typed male and female perceivers interact with male and female targets. In contrast to the Snyder et al. (1977) study, Andersen and Bem (1981) did not find that the male perceivers led female targets who they believed were attractive to respond in more pleasant and socially skilled ways.

Some allegedly attractive targets did respond more warmly than allegedly unattractive targets—but only when perceivers were sex-typed women. In contrast, androgynous female perceivers (those who described themselves as having both feminine and masculine characteristics) created a “boomerang” effect: Unattractive targets interacting with them were actually rated more favorably than were the attractive targets! Thus, the only attempt to replicate the classic Snyder et al. study (1977) almost completely failed: (1) There was no overall or general tendency for beliefs about one's partner's attractiveness to create a self-fulfilling prophecy, (2) male perceivers in particular did not evoke self-fulfilling behavior from their female interaction partners, (3) sex-typed female perceivers did evoke self-fulfilling behavior from their interaction partners, and (4) androgynous female perceivers actually evoked expectancy-disconfirming behavior from their partners.

A great story and a great “potential existence” demonstration. Snyder et al. (1977) is a great study, a classic. In the retelling (as in reviews), it makes for a great and dramatic story. And it provides an excellent experimental demonstration of the potential existence of self-fulfilling prophecies.

Of course, it is also only a single study. And like nearly all single studies, it had significant limitations. Given those limitations, and that the only attempt to replicate largely failed, it is not at all clear that this study provides a sound basis for any conclusions about the self-fulfilling effects of social beliefs in general or social stereotypes in particular. Despite the drama of the study, and the flare with which it can be described, this does not seem to provide much terra firma for concluding that expectancy effects are powerful and pervasive.

THE SEXIST INTERVIEWER STUDIES

The sexist interviewer studies (von Baeyer, Sherk, & Zanna, 1981; Zanna & Pack, 1975—see Chapter 4) showed that women were more likely to engage in traditional sex-role behaviors if they believed that an attractive man interviewing them for a job valued such behaviors. In some sense, then, these studies showed that attitudes toward sex roles could be self-fulfilling.

Limitations. These studies did not, however, show that male perceivers hold inaccurate gender-based expectations that, through self-fulfilling prophecies, they “impose” on unsuspecting women. In both studies, the expectations, values, or beliefs held by male interviewers were not examined, assessed, or manipulated. Indeed, in the Zanna and Pack (1975) study, there was no male interviewer at all (the women merely believed they would be interviewed)! Thus, neither of these studies showed that erroneous sex stereotypes held by male perceivers are self-fulfilling. They only showed college-age women slant their behavior in such a manner as to appeal to an attractive male interviewer. When the attractive male interviewer supposedly held traditional sex-role beliefs, this meant acting in a manner consistent with traditional sex roles; when the attractive male supposedly held nontraditional sex-role beliefs, this meant acting in a manner *inconsistent* with traditional sex roles.

A very narrow interpretation of this set of studies is clearly justified: Women (and probably men, too) sometimes slant their attitudes and behaviors in the direction of appealing to people from whom they want something. To conclude from this that sex stereotypes are, therefore, generally or powerfully self-fulfilling is to take a conceptual leap that goes well beyond the insights these studies did provide.

Replication. Although the sexist interviewer studies (von Baeyer et al., 1981; Zanna & Pack, 1975) have not been exactly replicated, there is clear evidence from both experimental and naturalistic studies that gender stereotypes can be self-fulfilling (Doyle, Hancock, & Kifer, 1972; Jacobs & Eccles, 1992; Palardy, 1969; Skrypnec & Snyder, 1982; see Jussim & Fleming, 1996, for a review). However, even these studies found typical expectancy effects (0.1 to 0.3) rather than particularly large ones; two of the studies showed that self-fulfilling prophecies advantaged girls over boys (Doyle et al., 1972; Palardy, 1969); one showed that self-fulfilling prophecies advantaged college males over college females (Skrypnec & Snyder, 1982); and one showed that self-fulfilling prophecies advantaged the child whose sex mothers believed was superior at math (which was more often, but not always, males—Jacobs & Eccles, 1992). Thus, although there is clear and convincing evidence that perceivers’ beliefs about males and females can be self-fulfilling, one would need to go well beyond the actual empirical evidence to conclude that self-fulfilling prophecies provide much more than a modest contribution to gender differences in power, wealth, status, and behavior.

Conclusion

The purpose of this chapter has not been to convince you that self-fulfilling prophecies never occur. This issue is not in doubt. They do occur. It has not even been to convince you that they occur less frequently and less powerfully than is often suggested or implied. My goal has been much more modest—to raise some doubts about the justification for believing that self-fulfilling prophecies are powerful or pervasive.

In virtually any social domain, I strongly suspect that nearly all readers can find many, perhaps countless, examples of both expectations that went unfulfilled (which cannot possibly be self-fulfilling prophecies) and expectations that were confirmed but that could not possibly have caused their own confirmation. History, sports, and economics all provide great examples of situations that reflect one of the main ideas of this book: Self-fulfilling prophecies do occur in daily life, but social reality usually influences expectations far more than expectations influence social reality.

I am, however, an empirical, research-oriented psychologist. Therefore, although common sense and personal experience can suggest possible insights into social phenomena, they are not, on their own, convincing. In my opinion, they should not be completely convincing to you either, because you and I might have opposite views of what constitutes common sense and because you and I might have completely opposing interpretations of some particular event. This is where hard, scientific research comes in. Therefore, this chapter also revisited several of the early classic self-fulfilling prophecy studies—studies frequently cited in support of claims emphasizing the power and pervasiveness of expectancy effects.

Despite their reputations, this critical review showed that *none* of these studies, either individually or when considered together, provided evidence that self-fulfilling prophecies were generally powerful or pervasive. One provided no evidence of self-fulfilling prophecy at all (Rist, 1970); one provided evidence that has been repeatedly challenged on methodological grounds, and that only constitutes evidence of typically modest self-fulfilling prophecies even if one takes its results at face value (Rosenthal & Jacobson, 1968a, 1968b); two did not address the self-fulfilling effects of perceivers' inaccurate expectations (von Baeyer et al, 1981; Zanna & Pack, 1975); and two have never been replicated (Snyder et al., 1977; Word et al., 1974). None showed that interpersonal expectations in general, or social stereotypes in particular, were typically inaccurate (because none assessed this).

Thus, this chapter was largely “negative” in the sense that it was not intended to argue what *does* happen. Instead, its purpose has been to raise doubts about the idea that self-fulfilling prophecies are powerful and pervasive. Although this single chapter cannot convey the whole picture regarding expectancy effects, it does provide another piece of the puzzle—neither daily life nor the classic studies provide compelling evidence that self-fulfilling prophecies are either typically large or particularly frequent. “So,” you should be wondering, “what does happen? Why aren’t self-fulfilling prophecies typically very powerful? How accurate or inaccurate is information-seeking, person perception, social stereotypes, and interpersonal expectations? Can we predict when expectancy effects are likely to be weak or nonexistent and when they might be more powerful?” These questions will be addressed over the next few chapters.

Notes

1. A discussion of why self-fulfilling prophecies may be particularly strong among military personnel is beyond the scope of this chapter. Given the overall weak evidence of self-fulfilling prophecies, many researchers, including myself, began seeking to identify conditions under which strong ones might exist—an endeavor I refer to as “the quest for the powerful self-fulfilling prophecy.” That quest will be described in a subsequent chapter.

With respect to the current chapter, however, although strong self-fulfilling prophecies in military contexts is an interesting and important phenomenon, in the context of the rest of the evidence presented in Table 6–1, I do not think it provides much *terra firma* for broad and general claims about the power and pervasiveness of expectancy effects. Even the grounds for the conclusion that “self-fulfilling prophecies are strong in military contexts” may be less than firm. Nearly all military studies included in McNatt’s (2000) meta-analysis focused on the Israeli Defense Forces (IDF) and came out of Dov Eden’s program of research (see McNatt, 2000, for the list of Eden’s studies). This is not to cast any aspersions on Eden’s work. However, psychology is filled with examples of individual researchers having a knack for demonstrating some phenomenon that proves difficult for other researchers to replicate. My point is only that the justification for drawing broad and general conclusions regarding the power of self-fulfilling prophecies even in military contexts, and, indeed, even in the narrow context of the IDF, would be considerably stronger if the same findings had been obtained by independent researchers.

2. Where does this 70% to 90% figure come from? On average, teams win 50% of their games. If the home team wins, on average, 55% of its games, then when away from home, they win, on average, 45% of their games. The home field advantage changes the outcome, on average, of 10% of a team’s games (the difference between home and away wins equals $55\% - 45\% = 10\%$). This means it does not change the outcome of 90% of the games. If the home team wins 65% of its games, then, when away from home, it wins 35% of its games. The home field advantage now changes the outcome of, on average, 30% of the games ($65\% - 35\% = 30\%$). This means that it does not change the outcome of 70% of the games.

3. A study by Chen and Bargh (1997) found that subliminally exposing White perceivers to an African American face led those perceivers (as compared to White perceivers subliminally exposed to a White face) to behave in a more hostile manner in a subsequent interaction with a White target and to evoke more hostile behavior from that target. Although interesting in its own right (indeed, it will be discussed at some length in Chapter 20), inasmuch as White perceivers never actually interacted with an African American target, I do not consider this study to be a replication. For those of you who cannot wait till Chapter 20, I will only point out here that the overall self-fulfilling prophecy effect in Chen and Bargh (1997) was about 0.2—right in the typical range of expectancy effects highlighted throughout this book. Thus, even if you interpret this study as replicating Word et al. (1974), like all the classics reviewed in this chapter, although it may provide credible evidence of the possible occurrence of self-fulfilling prophecies, it provides no evidence of particularly powerful ones.

4. Actually, this is not quite right. For the statistically inclined, it means that some linear combination of the 21 means differed between groups. For most practical purposes, however, this is equivalent to concluding that at least one mean differed between groups.

7 You Better Change Your Expectations Because I Will Not Change (Much) to Fit Your Expectations

SELF-VERIFICATION AS A LIMIT
TO SELF-FULFILLING PROPHECIES

My Fourth Grade Spelling Experience

I was a sickly kid through most of elementary school. I missed almost 80 days of school in first grade and over 60 in second grade. I have warm and fuzzy childhood memories of being home from school and having my mom spend all this time with me trying to help stay caught up in my school work, doing flash cards for spelling and simple math problems. One time, I had both the measles and croup at the same time. I was home sick in bed and was having a hard time breathing. The next thing I knew, I woke up with an oxygen mask and surrounded by what I thought were police (they were probably emergency medical personnel). I was then whisked into an ambulance, which sirened through the streets of Brooklyn to rush me to a hospital. I felt much better after the oxygen mask and thoroughly enjoyed ripping through red lights in the siren-blaring, speeding ambulance.

All this illness eventually took a toll on my school performance. My elementary school used tracking, also known as ability grouping. The “smartest” kids went into one class, the next group in another, and so on. Because I had missed so much in first and second grade, they put me one class from the bottom in third grade. But by third grade, my health began to improve. I was still pretty sickly compared to most kids—I missed over 20 days of school—but nowhere near as bad as in first and second grade. So I did pretty well in third grade. Well enough for them to put me in the top class in fourth grade.

My fourth grade teacher, Mr. Sunshine (this really was his name—he was a great teacher by the way) used within-class grouping for spelling. The best spellers were placed into the top group and had to learn how to spell difficult words—typically on a fifth- to seventh grade level. The other group received fourth to fifth grade words. Since I came up off the near bottom in third grade, Mr. Sunshine naturally placed me in the lower group. Although I liked Mr. Sunshine quite a lot, this annoyed me. I knew I could learn those tougher words. But Mr. Sunshine had a policy that gave me cause for hope—anyone who got three 100s in a row on spelling tests would automatically get bumped up to the higher group.

I was sure I could do it. My problem was (as most people who know me well can testify is still true today) that I often make careless or thoughtless mistakes. So I never could quite get those three 100s in a row. I might get a 100, then a 90, then a 95, then two 100s in a row, then another 95, and so on. But I knew I should have been in the higher group, so every time I fell short, it made me more determined to get my three 100s in a row.

But I never did. Why? Because Mr. Sunshine eventually realized that, even without three 100s in a row, the lower group was too easy for me—and he moved me up. And I continued to get 90s and 100s on my spelling tests (OK, I admit it, I had an 80 and 85 once in a while).

Hey, what happened to self-fulfilling prophecy? Ability grouping is often depicted as an unmitigated evil serving to create and maintain a caste-like system in which the academically rich get richer and the academically poor get poorer (e.g., Oakes, 1985; Rist, 1970). If so, they should be a great mechanism for creating self-fulfilling prophecies—just relegate low-expectancy kids to the low classes and put the supposedly smart kids in the top classes, and tracking will ensure the fulfillment of those expectations. I should have become a mediocre speller, right? Somehow, though, I managed to escape the near-bottom classes. Even so, why didn't Mr. Sunshine's early-year low expectations fulfill themselves? Why didn't my spelling skills decline, or at least improve too slowly to warrant moving up?

As I have been trying to point out throughout this book (except for Chapters 4 and 5), interpersonal expectations are often just not very powerful. Chapters 6 through 9 identify a slew of reasons why. One such reason is self-verification.

What Is Self-Verification?

Self-verification refers to the idea that people are often highly motivated to see themselves in a manner consistent with their own long-standing and deep-seated self-views (Swann, 1987). For example, people are more highly motivated to seek out information that confirms their self-perceptions than disconfirms their self-perceptions (Swann & Read, 1981a, 1981b). If people like themselves, they spend more time finding out what a person who also likes them thinks about them than what a person who dislikes them thinks about them; people who dislike themselves, however, spend more time finding out what a person who dislikes them thinks about them (Swann & Read, 1981b).

People with high self-esteem often consider positive feedback on some performance much more accurate and believable than negative feedback; people with low self-esteem, however, often consider negative feedback on some performance more accurate and believable than positive feedback (Jussim, Yen, & Aiello, 1995; Swann, Griffin, Predmore, & Gaines, 1987).

And people with high self-esteem often interpret feedback (both positive and negative) more positively than do people with low self-esteem (Jussim, Coleman, & Nassau, 1987; Jussim, Yen, et al., 1995).

Self-Verification as a Limitation to Self-Fulfilling Prophecies

What does this have to do with self-fulfilling prophecies? Apparently, quite a lot. A strong self-concept, it seems, constitutes the psychological rudder that assists people in finding their own way through the potentially stormy seas of others' expectations. At least, that seems to be the message from a series of studies that have pitted self-verification against self-fulfilling prophecies. Self-verification, apparently, extends beyond the motivation to see one's actions and achievements in a manner consistent with one's self-perceptions. As the following studies all show, it also includes *convincing other people* to view one much as one views oneself.

Swann and Ely (1984). Swann and Ely (1984) performed the first study that simultaneously examined self-verification and self-fulfilling prophecy. Their paper was titled "A Battle of Wills" to capture the idea that, when perceivers' expectations conflicted with targets' self-perceptions, self-fulfilling prophecy and self-verification were in direct opposition to one another. Would perceivers impose their expectations on targets and lead targets to confirm them? Or would targets convince perceivers to change their expectations?

Before examining these questions, however, Swann and Ely (1984) reasoned that both perceiver and target certainty (regarding their expectations and self-perceptions, respectively) would likely influence the outcome. In general, they suggested that certainty (on both the perceiver's and target's part) would increase the likelihood of "winning" the battle. Perceivers who were more certain of their expectations should be more likely to produce self-fulfilling prophecies; targets who were more certain of their self-perceptions should be more likely to self-verify.

What would happen when both perceivers and targets were high in certainty? Swann and Ely (1984) predicted that targets would "win" such battles. Targets' beliefs about themselves would likely be held more confidently than would perceivers' beliefs about targets (targets have a much more vast amount of personal experience to support such beliefs than would even the most certain perceivers).

To compare self-verification with self-fulfilling prophecy, Swann and Ely (1984) had college women interview other college women. The interviewer was the perceiver; the interviewee was the target. Swann and Ely (1984) identified targets who considered themselves to be either introverted or extroverted, and who were either high or low in the certainty with which they held such beliefs. Prior to the interview, they led perceivers to develop *opposing* (or erroneous) expectations for these targets. Interviewers of introverts were led to believe interviewees were extroverts; interviewers of extroverts were led to believe interviewees were introverts. Perceiver certainty was manipulated by the consistency of the evidence they received—low certainty was induced by providing mixed evidence of target introversion or extroversion; high certainty was induced by providing consistent evidence of target introversion or extroversion. Table 7-1 presents an overview of the experimental design and predictions of this study.

TABLE 7-1

Summary of Design and Predictions from Swann and Ely (1984)				
	Low Perceiver Certainty		High Perceiver Certainty	
	Low Target Certainty	High Target Certainty	Low Target Certainty	High Target Certainty
Introvert expectation/ extrovert self-view	Self-verification	Self-verification	Self-fulfilling prophecy	Self-verification
Extrovert expectation/ introvert self-view	Self-verification	Self-verification	Self-fulfilling prophecy	Self-verification

Notes. In this study:

1. Self-verification means that the perceiver's beliefs and behaviors regarding the target changed to become more consistent over time with the target's self-perceptions.
2. Self-fulfilling prophecy means that the target's self-perceptions changed to become more consistent with the perceiver's expectation over time.

The results of the study generally confirmed these predictions (see text for details).

Swann and Ely (1984) reasoned that changing a perceiver's expectation through self-verification might take some time, and that changing a target's self-perception through self-fulfilling prophecy might take some time. In order to allow these processes ample opportunity to unfold, *three* interview sessions were conducted. After all sessions were completed, undergraduate judges rated the degree of extroversion of the targets from tape recordings of the interview sessions.

Changes in perceiver expectations were not assessed directly. They were assessed indirectly, by evaluating the types of questions the perceiver/interviewers asked the targets. For example, "Do you like to go to big parties?" was a question more likely to be asked of someone believed to be extroverted. "Do you have trouble meeting people and making friends?" was a question more likely to be asked of someone believed to be introverted. At each session, the experimenter provided the perceiver/interviewer with a list of 12 questions from which she could select 5 to ask the target. It was a new list consisting of a different set of 12 questions at each session, but each list included 6 extroverted questions and 6 introverted questions. Thus, by assessing the number of extroverted and introverted questions perceivers asked, Swann and Ely (1984) could identify their (implicit) expectation.

Results, Session 1. Perceivers asked more expectancy-consistent questions, but only when they held their expectation with high certainty. There was no reliable tendency for perceivers who were not so certain of their expectation to ask expectancy-consistent questions.

What about the target's behavior? Did it change to more closely correspond to perceivers' expectations? When the target was certain of her self-conception, her behavior did not change at all. Extroverts acted in a more extroverted manner than introverts, regardless of whether perceivers held their expectations with high or low certainty. Even when the target was low in certainty herself, there was no self-fulfilling prophecy *if* the perceiver was also low in certainty. Extroverts still acted more extroverted; introverts still acted more introverted.

A self-fulfilling prophecy occurred only when the target was low in certainty and the perceiver was high in certainty. In that situation, introverts became somewhat more extroverted

and extroverts became somewhat more introverted so that judges rated their degree of extroversion as virtually identical.

Results, Session 2. Most perceivers in Session 1 received disconfirmatory responses from targets. Did perceivers rigidly resist changing their expectations in the face of this disconfirmation? Or did their expectations (as reflected in the types of questions they asked) change? Their expectations changed. Low-certainty perceivers faced with high-certainty targets changed the most. This would seem pretty obvious, inasmuch as these perceivers did not hold their expectations particularly strongly to start with, and they received loud and clear disconfirming evidence from targets in the first session. Indeed, these low-certainty perceivers showed a *reversal* of their expectations—perceivers given the introverted expectation (facing extroverted targets) asked more *extroverted* questions in Session 2; perceivers given the extroverted expectation (facing introverted targets) asked more *introverted* questions in Session 2. This looks a lot like accuracy rather than bias, in that perceivers' expectations seemed to reflect, rather than cause, targets' behavior.

Although changes in perceiver expectations were strongest when perceivers were low in certainty and targets were high in certainty, such changes occurred across the board. In Session 2, low-certainty perceivers asked more *expectancy-disconfirming* than expectancy-confirming questions even when targets were low in certainty (although this reversal was not as pronounced as when targets were high in certainty). Even among high-certainty perceivers, the tendency to ask expectancy-consistent questions was greatly reduced. There was no reliable difference in the types of questions asked by high-certainty perceivers, regardless of whether they held introverted or extroverted expectations.

What about target behavior? It was very similar to the pattern of Session 1. High-certainty targets acted in line with their self-conceptions, not with perceiver expectations, as did low-certainty targets interviewed by low-certainty perceivers. Again, however, there was no reliable difference between low-certainty introverts and extroverts when they were interviewed by a high-certainty perceiver, indicating a continued self-fulfilling prophecy effect.

Results, Session 3. As Swann and Ely (1984, p. 295) put it: "In light of the steadfast refusal of most targets to provide perceivers with expectancy-consistent evidence in the first two sessions, we suspected that perceivers would abandon efforts to elicit such information in Session 3. This was the case; the questions perceivers asked in Session 3 did not differ as a function of their expectancies, whether targets were high or low in certainty. . . ."

Results for targets mirrored Sessions 1 and 2. High-certainty targets acted in line with their self-perceptions and were not affected by perceiver expectations, as did low-certainty targets interacting low-certainty perceivers. Judges saw no extroversion/introversion differences in the behavior of low-certainty targets interacting with high-certainty perceivers, again providing some evidence of self-fulfilling prophecy.

Other results. At the end of all three sessions, Swann and Ely (1984) performed a final series of analyses. These analyses showed that:

1. Low-certainty perceivers completely revised their beliefs about targets, coming to believe, appropriately, that self-perceived extroverts were more extroverted than self-perceived introverts.
2. High-certainty perceivers also heavily revised their expectations. These perceivers ended up rating introverts and extroverts similarly. Although this may appear to

indicate that high-certainty perceivers did not change their expectations as much as did low-certainty perceivers, there is another interpretation. Perhaps they not only were more certain of their expectations—but also held *stronger* expectations. Perhaps the consistent evidence of targets' introversion/extroversion led them to believe that targets were *more* introverted or extroverted than low-certainty perceivers believed. If so, then rating introverts and extroverts similarly at the end may reflect as much *change* for high-certainty perceivers as was found among low-certainty perceivers who viewed extroverts as more extroverted than introverts at the end. Swann and Ely (1984) did not provide the data necessary for choosing among these two alternative interpretations.

3. Perceivers' expectations did not significantly change targets' self-conceptions.
4. The correlation between the type of questions the perceivers asked and judges' ratings of the targets was near zero. This means that asking people supposedly introverted expectancy and extroverted expectancy questions had little effect on their behavior. (This point will be extremely important in the *next* chapter, which addresses the extent to which expectations bias social information gathering.)
5. Targets' behavior was highly consistent from session to session, whereas perceivers' question-asking patterns changed quite a lot. That is, perceivers' behavior changed much more than did targets' behavior. This suggests that (a) perceivers' expectations were highly flexible and responsive to disconfirming evidence and (b) targets' behavior was not readily pushed around by perceivers' erroneous expectations.

Subsequent Research on Self-Verification Versus Self-Fulfilling Prophecy

Major, Cozzarelli, Testa, and McFarlin (1988). This was the first attempted replication of the Swann and Ely (1984) study and was similar in that previously unacquainted pairs of college students interacted with one another; one member of each pair was given a false expectation for the other; and judges rated the behavior of the perceivers and targets. This research also differed in some important respects: All subjects were males; there was only a single, 10-minute interaction (rather than three sessions); the interaction was not face to face (they used headphones and microphones); and the expectancy involved sociability, rather than intro-/extroversion. In addition, they also examined whether target self-consciousness influenced degree of self-verification and self-fulfilling prophecy.

Major et al. (1988) concluded that their findings were more consistent with an expectancy confirmation perspective than with self-verification. In my view, however, such a conclusion was not justified by their results. Next, therefore, I describe their results in some detail.

Did perceivers' expectations influence targets' behavior and create a self-fulfilling prophecy? To evaluate this question, the outside judges rated the sociability of the targets at the beginning and end of the interaction. The ratings of the high-sociability (low expected sociability) targets declined substantially from beginning to end, whereas the ratings of the low-sociability (high expected sociability) targets declined only slightly. Although the reason why all targets' sociability ratings declined is unclear, the ratings of the high-sociability (low expected sociability) targets changed in the direction of perceivers' expectations. Thus, for high-sociability targets, there was evidence of self-fulfilling prophecy.

Did perceiver expectations change target self-conceptions (ala self-fulfilling prophecy), or did targets resist perceivers' expectations (ala self-verification)? There were four cells in the study: high and low target self-consciousness by high- and low-sociability expectations (which were always opposite targets' self-perceived sociability). Supporting self-verification, in three of the four cells, there was minimal change in targets' self-perceived sociability. The fourth cell showed evidence of self-fulfilling prophecy. Only highly self-conscious targets who perceived themselves to be low in sociability *and* who were believed to be highly sociable showed substantial evidence of change—they saw themselves as considerably more sociable at the end of the interaction.

Did targets change perceivers' expectations? Major et al. (1988, p. 355) concluded that their results were "... consistent with expectancy-confirmation predictions" and that "... in contrast to Swann & Ely's (1984) study, perceivers tended to cling to their false beliefs about targets' sociability." This is because both before and after the interaction, perceivers given a high-sociability expectation still rated the targets as more sociable than did perceivers given a low-sociability expectation.

Although this is certainly true, I do not think their obtained pattern justifies a conclusion that perceivers "clung" to their expectations. Their results showed that:

1. Perceivers' expectations for high sociables (low expected sociables) changed dramatically over the course of the interaction. These perceivers came to view their interaction partners as *much* more sociable than they first thought.
2. Perceivers' expectations for low sociables (high expected sociables) changed slightly over the course of the interaction. Perceivers came to view their interaction partners as slightly less sociable than they first thought.¹

The term "cling" has connotations that were not made explicit in the article—it seems to imply something like a motivated determination on the part of perceivers to not have their expectations disconfirmed; it seems to imply something like "rigidly resistant to disconfirming evidence." Yet, at least for low-sociability-expectation perceivers, there was not a shred of evidence of rigidity or clinging—they changed their expectations quite dramatically after a mere 10-minute interaction! I suspect that this may be because it is pretty hard to misinterpret highly sociable behavior. People who see themselves as highly sociable, if they are correct, act in friendly, warm, outgoing ways. If so, it is clear that perceivers did not miss it.

That there was even slight change among perceiver beliefs regarding low sociables could also be viewed as impressive after a mere 10-minute interaction. Especially if low sociables tended to be quieter and less outgoing, their behavior could easily have been seen as less clearly disconfirming than was the behavior of high sociables. Perhaps they were really highly sociable with their friends and acquaintances, but not with strangers in a lab, perhaps they just happened to be tired, etc. Relatively quiet behavior could be viewed as providing less disconfirming evidence regarding a highly sociable expectation than does gregarious behavior (especially in an interaction among strangers) of a low-sociability expectation.

Overall, therefore, the conclusions of Major et al. (1988) emphasizing expectancy confirmation notwithstanding, as far as I can tell, their results provided at least as much evidence of self-verification as self-fulfilling prophecy. Table 7–2 summarizes their results. Major et al.

TABLE 7.2

Summary of Results from Major et al. (1988)

OUTCOMES	EVIDENCE OF EXPECTANCY- CONFIRMATION	EVIDENCE OF SELF-VERIFICATION	AND THE WINNER IS:
Target Behavior	Low expected sociables (high self-perceived sociables) became less sociable (self-fulfilling prophecy)	High expected sociables (low self-perceived sociables) became slightly less sociable (weak self-fulfilling prophecy)	Self-Fulfilling Prophecy
Target Self- Concept	High expected sociables (low self-perceived sociables) high in self-consciousness increased self-ratings of sociability. (self- fulfilling prophecy in one condition)	All three other conditions showed little change in self-concept (self- verification in three conditions)	Self-Verification
Perceiver Expectations	Perceiver ratings of high expected sociables (low self-perceived sociables) changed little (expectancy confirmation)	Perceivers rated low expected sociables (high self-perceived sociables) much more sociable at the end of the interaction (self-verification)	Tie

(1988) found behavioral evidence of self-fulfilling prophecy among half the targets (the high sociables), self-concept evidence of self-fulfilling prophecy among one-quarter of their targets (low sociables high in self-consciousness), and self-concept evidence of self-verification among the other three-quarters, and they found that even after a single 10-minute interaction, perceivers' expectations changed to become more consistent with targets' self-conceptions (although this was much more true among those interacting with high self-perceived sociables than with low self-perceived sociables). This is clearly a mixed pattern of self-fulfilling prophecy and self-verification.

McNulty and Swann (1994). McNulty and Swann (1994) were the first to examine the "battle of wills" outside of the laboratory—they performed two studies investigating self-fulfilling prophecy and self-verification among previously unacquainted college roommates over 10 weeks. They assessed each roommate's (1) perceptions of the other roommate's characteristics (they called these "appraisals") and (2) self-perceptions. Their first study examined ratings regarding 10 attributes: academic ability, social skill, athletic ability, attractiveness,

neuroticism, extroversion, openness to new experiences, agreeableness, conscientiousness, and global self-esteem. Each of these ratings was obtained twice (2nd and 12th week of the semester) from 69 pairs of roommates.

The self-fulfilling prophecy hypothesis is that roommate appraisals early in the semester would predict changes in target self-concept by the end of the semester. There was statistically significant evidence of such change on 5 of the 10 ratings (social skill, athletic ability, extroversion, openness, and conscientiousness—effect sizes ranging from .14 to .23 on these five). The self-verification hypothesis predicts that self-perceptions early in the semester would predict changes in roommate appraisals by the end of the semester. There was statistically significant evidence of self-verification effects on 4 of the 10 ratings (academic ability, social skill, athletic ability, and extroversion—effect sizes ranged from 0.25 to 0.37 on these four).

In addition, McNulty and Swann (1994) identified dyads among whom a self-fulfilling prophecy had occurred, among whom self-verification occurred, and among whom both occurred. Depending on the particular rating, a self-fulfilling prophecy occurred among 23% to 35% of all dyads; self-verification occurred among 20% to 41% of all dyads.²

Their second study was highly similar and yielded largely similar results. In this second study, however, there were 95 pairs of roommates, but McNulty and Swann (1994) only examined five ratings (academic ability, social skill, athletic ability, attractiveness, and self-esteem). This study yielded statistically significant evidence of self-fulfilling prophecy on only one measure (academic ability, effect size = 0.18), and statistically significant evidence of self-verification on two measures (academic and athletic ability; effect sizes = 0.15 and 0.20, respectively). Self-fulfilling prophecy and self-verification each occurred among about 25% of dyads.

Swann, Milton, and Polizer (2000). This study examined the role of self-verification and self-fulfilling prophecies (again, called “appraisal effects” by the authors) in small working groups of four to six new MBA students. They assessed 11 attributes, such as academic, leadership, and social ability. Group expectations (“appraisals”) were operationalized as *group* perceptions of the target, rather than as individuals’ perceptions of one another.³ Group expectations and self-perceptions were assessed at three points over a single semester (the first assessment was at the very beginning of the semester, before the groups were formed; Time 2 was 9 weeks into the semester; and Time 3 was at the end of the 15-week semester).

Self-fulfilling prophecy. The first set of analyses assessed self-fulfilling prophecies.

Did Time 1 group expectations change self-perceptions? To address this question, Swann et al. (2000) used Time 1 group expectations to predict Time 2 self-perceptions controlling for Time 1 self-perceptions. Results showed statistically significant evidence of self-fulfilling prophecy on 5 of the 11 variables. These five effects ranged from 0.09 to 0.15.

Self-verification. The next set of analyses addressed self-verification. Did Time 1 self-perceptions change group expectations? To address this question, Swann et al. (2000) used Time 1 self-perceptions to predict Time 2 group expectations, controlling for Time 1 group expectations. There was statistically significant evidence of self-verification on 8 of the 11 variables. These eight effects ranged from 0.10 to 0.31.

Total self-fulfilling prophecy versus self-verification effects. Swann et al. (2000) also estimated the total amount of self-fulfilling prophecy and self-verification for each target by assessing whether (1) later self-perceptions moved closer toward initial group expectations and (2)

later group expectations moved closer toward initial self-perceptions. These analyses showed that self-fulfilling prophecy effects occurred among 25% to 38% of all targets and that self-verification effects occurred among 44% to 48% of all targets.⁴

Madon et al. (2001). We (I am one of the “et al.”) examined the battle of wills in the “original” self-fulfilling prophecy context—the classroom. Relations between teacher expectations and students’ self-concept of ability were examined in 108 sixth grade classes, including nearly 1,700 students. Teacher expectations and students’ self-concepts were assessed at two different time points—early and late in the school year. In addition, all analyses controlled for a slew of potential influences on both teacher expectations and student self-concept (prior grades and standardized test scores, several motivational variables, and students’ demographic background). This was important because there could be a great deal of overlap between teacher expectations and student self-concept, not because of *either* self-fulfilling prophecy or self-verification, but because of accuracy (even if expectations and self-concept have no causal influence on one another, they might still correlate highly if they are both based on, for example, students’ prior academic achievement).

Results showed evidence of both small self-fulfilling prophecies and small self-verification effects. Consistent with the self-fulfilling prophecy prediction, teacher expectations early in the year predicted slight changes in student self-concepts by the end of the year (effect size = 0.12). Consistent with the self-verification prediction, student self-concept early in the year predicted slight changes in teacher expectations by the end of the year (effect size = 0.04). Although small, both effects were statistically significant (thanks to the large sample). These results might appear to indicate that self-fulfilling prophecy was stronger than self-verification, but an additional analysis showed that there was no significant difference between these two effect sizes. Thus, neither effect was very large and they were statistically no different from one another.

SELF-VERIFICATION AND SELF-FULFILLING PROPHECY: CONCLUSIONS

Research on self-verification highlighted limitations to self-fulfilling prophecies in a host of ways:

1. Swann and Ely (1984), McNulty and Swann (1994), and Swann et al. (2000) showed that self-fulfilling prophecies only occurred about a quarter to a third of the time (in two of eight cells, among about 25% of dyads, and among 25% to 38% of targets, respectively). Major et al. (1988) showed that self-fulfilling prophecies occurred in half the cells with respect to behavior and in 25% with respect to target self-concept. Our study (Madon et al., 2000) did not report this type of data.
2. The studies reporting effect sizes (Madon et al., 2001; McNulty & Swann, 1994; Swann et al., 2000) showed that self-fulfilling prophecy effects ranged from 0 to a high of about 0.2.
3. Two studies showed that, rather than rigidly clinging to their expectations, perceivers generally changed their expectations in response to disconfirming targets (Major et al., 1988; Swann & Ely, 1984). The other studies did not examine this issue.

Bottom line: Self-verification is one reason self-fulfilling prophecies are not typically powerful or pervasive. People are not rudderless ships on the seas of others' expectations. The self is a rudder, and a pretty powerful one at that.

Because of Swann and Ely's (1984) "battle" framing, there was, perhaps, a greater sensitivity to "keeping score" than in much prior social psychological research on expectancies. This seems to have led to noticeably greater attention being paid to details such as the actual size of the expectancy effect and whether or not perceivers' expectations were rigidly resistant to change and whether they dramatically biased interpretation of target behavior. Such attention clearly uncovered the pattern that I have been emphasizing and will continue to emphasize throughout this book: (1) Self-fulfilling prophecies are real in the sense that people's expectations do sometimes influence others' behavior, self-perceptions, and characteristics; but (2) such effects, rather than being powerful and pervasive, are typically small and fleeting; and (3) although perceivers' expectations do sometimes lead to biases and errors, they are often reasonably accurate and highly responsive to disconfirming information, rather than rigidly resistant to change.

Chapters 3, 4, 6, and this one have addressed points 1 and 2 regarding self-fulfilling prophecies. Point 3 is really two points: (1) Perceiver expectations are often reasonably accurate (an issue addressed in detail starting in Chapter 10) and (2) although expectations can bias perception, they are typically highly responsive to disconfirming information. This latter point is addressed in depth in Chapters 8, 9, and 18.

Notes

1. For the statistically inclined, Major et al. (1988) did not test whether either of these changes in perceivers' expectations from pre- to postinteraction were statistically significant. However, the big change (perceivers' expectations for low sociables [high expected sociables]) would almost certainly have been statistically significant, because it is larger than the statistically significant difference in postinteraction expectations that they did report. Whether the smaller change among perceiver expectations for high sociables (low expected sociables) would also be statistically significant cannot be determined from the data that they reported.

2. Both here and in the Swann et al. (2000) study described next, these numbers are purely descriptive, in the sense that it was not possible to perform statistical tests. A person was classified as experiencing a self-fulfilling prophecy effect if the T2 self-perception was closer to the T1 expectation than was the T1 self-perception (i.e., the T2 self-perception moved closer to the T1 expectation). Similarly, they were classified as experiencing a self-verification effect if the T2 expectation was closer to the T1 self-perception than was the T1 expectation (i.e., the T2 expectation moved closer to the T1 self-perception). Although heuristically useful, this means that *any movement, no matter how small*, was classified as evidence of self-verification or self-fulfilling prophecy. This is still useful for comparing extent of self-verification versus self-fulfilling prophecy, but it also means that the effect sizes obtained from the regression analyses are probably the best indicators of the overall power of both self-fulfilling prophecy and self-verification effects.

3. For the statistically inclined, they did this by averaging all group members' expectations (appraisals) of each target (so there would be a separate expectation [appraisal] variable for each

target), after removing target effects (see Kenny, 1994). All analyses then employed hierarchical linear modeling (a sophisticated statistical procedure) to assess relations between expectations and self-perceptions, while controlling for group-level effects.

4. See footnote 2 regarding interpretation of these results. Also, at Time 3, Swann et al. (2000) did not collect expectation data, so it was impossible to test for self-verification effects on Time 3 expectations. However, they did assess the self-perception variables and did test for self-fulfilling prophecies, finding essentially the same pattern as found at Time 2—significant evidence of self-fulfilling prophecy on 4 of 11 variables, with effects on those variables ranging from 0.09 to 0.14. Swann et al. (2000) also examined the extent to which self-fulfilling prophecy and self-verification effects predicted group performance outcomes, but this issue, though interesting, is beyond the scope of this chapter.

8 The Less Than Awesome Power of Expectations to Distort Information-Seeking

“WHEN DID YOU stop beating your wife?” This is the classic example of a leading question. It is so well-known and obvious that it has become trite. Well, what about “What would you do to liven up a party?” Not so well-known. Ever ask anybody that question? I haven’t. Ever been asked? Me neither. Although there are probably some exceptions, such nakedly obvious, biased, leading questions rarely seem to come up—at least not in my daily experience. Someone might ask me, “Hey, how is your methods class going?” but I have never been asked, “What torture have you cooked up for those students this week?”

Chapter 5 described Snyder and Swann’s (1978) research showing that, when asked to test a hypothesis about a stranger in an interview, people select interview questions that are biased to the point of virtually guaranteeing responses that confirm the hypothesis. One example was asking people believed to be extroverts, “What would you do to liven up a party?” Probably not even the most shy and withdrawn introverts could provide a disconfirming answer to this question. “Well, maybe I would start playing some loud, fast music” or “I would break out the wine and beer” or virtually any reasonable answer would lead the interviewee to sound like an extrovert, thereby seeming to confirm the hypothesis. To answer, “Well, in point of fact, I probably would not do anything, and, indeed, I am uncomfortable in such situations so that it would be unusual for me to find myself at any sort of party, even a dull one” would be hypothetically possible, but it would be such an odd, awkward response that the probability of anyone providing such a response is vanishingly small.

The thing is, except perhaps in a heavily watered-down form (discussed toward the end of this chapter), this pattern of people selecting questions is virtually guaranteed to evoke an expectancy-confirming response has not held up. Here, as all throughout Chapters 6

through 9, I will *not* be claiming that expectations have absolutely no influence on how people go about seeking information, or that lay information-seeking is absolutely and completely rational and scientific. I will be suggesting, however, that the accumulated social psychological evidence shows that (1) people have virtually no tendency to spontaneously ask the type of highly constraining questions from which participants were required to choose in the Snyder and Swann (1978) study, (2) intuitive social information-seeking is mostly (though not 100% completely) dominated by an even-handed or balanced search for confirming and disconfirming information, although (3) there is a slight tendency to seek or prefer confirming information, which, it turns out, might sometimes lead to small self-fulfilling prophecies.

Do People Naturally Ask Highly Constraining Questions?

No. Not at all. I am very uncomfortable with absolutes like “never” or “always,” so let’s just say “almost” never. The Snyder and Swann (1978) study has long been cited as showing that perceivers constrain targets’ reactions in such a manner as to almost ensure expectancy confirmation (e.g., Fiske & Taylor, 1991; Gilbert, 1995; Markus & Zajonc, 1985; Sommers & Norton, 2008). Such a claim is far too strong; it was not justified by the Snyder and Swann (1978) research, and subsequent research has resoundingly disconfirmed this claim.

Snyder and Swann (1978) constrained perceivers to ask constraining questions. One of the most obvious limitations to Snyder and Swann’s (1978) study was that they *required* participants to ask highly constraining questions. They did not examine the types of questions that people spontaneously develop to test their hypotheses or expectations. Snyder and Swann (1978) attempted to be fair in that there were two opposite types of highly constraining questions: (1) questions to which answers would confirm the hypothesis and (2) questions to which answers would confirm the opposite hypothesis (e.g., “What things do you dislike about loud parties?”). But both types heavily constrain the type of answer likely to be received.

Contrast these types of questions, with, for example, “Do you try to liven up dull parties?” or “Do you dislike loud parties?” A person could easily answer “yes” or “no” to either one. Disconfirming responses are neither awkward nor uncomfortable. These questions are not constraining at all.

Thus, a fair and accurate characterization of Snyder and Swann’s (1978) findings would be something like: “When perceivers are required to choose questions from a list containing only questions that constrain the answer to confirm or disconfirm their expectations for a target, perceivers prefer questions that constrain the answer to confirm, rather than disconfirm, their expectations.” This is a far more complex, narrow, and nuanced claim than, for example, that “People seek to confirm their hypotheses” or “People ask questions in such a manner as to almost guarantee that their hypotheses will be confirmed” or “Seeking confirmation biases information gathering.” The above quotes are mine, but they capture the spirit of how this research has often been interpreted (e.g., Gilbert, 1995; Hamilton, Sherman, & Ruvalo, 1990; Markus & Zajonc, 1985; Sommers & Norton, 2008).

OK, so because of their methodological limitations, Snyder and Swann (1978) did not show that information-seeking is heavily biased toward expectancy confirmation. Even if you

agree that their research did not justify this conclusion, it does not mean their conclusions were necessarily wrong. Perhaps research that addressed or eliminated their methodological limitations would show that people have a powerful tendency to seek expectancy-confirming information. So let's next examine two sets of studies that followed on their heels.

Do people spontaneously ask constraining questions? This was precisely the question addressed by Trope, Bassok, and Alon (1984). Participants in their study were led to believe that the researchers were investigating how people learn about others' personality. To this end, participants were first asked to develop questions for an upcoming interview, and second, to conduct the interview. Half were asked to assess whether the interviewee was an introvert (the "introvert hypothesis condition"); half were asked to assess whether the interviewee was an extrovert (the "extrovert hypothesis condition"). They then read a brief excerpt from Snyder and Swann's (1978) procedures that described the characteristics of introverts and extroverts.

Coders categorized the interviewers' questions into one of six groups. *Biased* questions referred to questions like those used by Snyder and Swann (1978)—questions that all but guaranteed that the interviewee would provide a confirming response. There were, of course, two types of biased questions—biased introvert questions and biased extrovert questions.

There were four types of unbiased questions. One type Trope et al. (1984) referred to as "consistent" questions. Although these questions were not constraining, a "yes" response would confirm the hypothesis. "Are you usually the initiator in forming new friendships?" is an example of an unbiased, extrovert-consistent question. "Do you usually go to movies alone" is an example of an unbiased, introvert-consistent question. A person could easily answer "no" to either of these without coming across as awkward or pugnacious. Nonetheless, in both cases, a "yes" response confirms the hypothesis. Thus, there were two types of consistent questions—extrovert consistent and introvert consistent.

There were two other types of unbiased questions. Trope et al. (1984) referred to questions that presented an introvert-consistent choice and an extrovert-consistent choice as "bidirectional" (e.g., "Do you prefer big or small parties?"). There were also "open-ended" questions—which probed for introversion/extroversion without an explicit choice and to which a "yes" or "no" response would not be appropriate (e.g., "How do you spend your Friday nights?").

What did they find? They performed two studies. Across the two studies, nearly 600 questions were generated, and a grand total of two—that's right, two—constrained interviewees' responses to confirm the hypothesis. Furthermore, about two-thirds of all questions were either bidirectional or open ended. Only a minority fell into the "consistent" category.

Furthermore, when focusing exclusively on the "consistent" questions, the introvert versus extrovert hypothesis (expectancy) made no difference. People with an introvert hypothesis were no more likely to generate introvert-consistent questions than were people with an extrovert hypothesis. Similarly, people with an extrovert hypothesis were no more likely to generate extrovert-consistent questions than were people with an introvert hypothesis.

Swann and Giuliano (1987) performed a study much like the one by Trope et al. (1984) and found that their question askers almost never generated questions anywhere near as constraining as those used in the Snyder and Swann (1978) study. Unlike Trope et al. (1984), however, Swann and Giuliano found that people did ask more expectancy-consistent questions (questions to which a "yes" answer would confirm the hypothesis) than

expectancy-inconsistent questions (questions to which a “yes” response would disconfirm the hypothesis).

Swann and Giuliano (1987) referred to these as “confirmatory questions.” This terminology was, in my view, unfortunate because it seems to have conveyed the idea that people naturally engaged in the type of biased information-seeking described by Snyder and Swann (1978). Of course, however, asking *if* people try to liven up a dull party is a very different (less biased, less constraining, far less “confirmatory”) type of question than asking “What would you do to liven up a party?” And, indeed, when this study is cited, it is often cited in support of the claim that people engage in expectancy-confirming information-seeking (e.g., Fiske & Taylor, 1991).

Again, such a claim goes too far. The study can be cited in support of the claim that “People prefer to ask questions to which a ‘yes’ response confirms their expectation.” This technical, narrower, and more nuanced conclusion is justified by their study. But it does not mean that people engaged in a search for information that was biased toward confirming their expectations.

Confirmation Versus Diagnosis in Information-Seeking

So maybe perceivers’ information-seeking strategies are not so distorted as to virtually force targets to provide responses that confirm perceivers’ expectations. The suggestion that they do so was a pretty extreme claim; but denying this extreme claim does not mean that people are perfectly objective information seekers, either. Even if people do not force targets’ responses to confirm their expectations, perhaps their behavior and strategies are sufficiently flawed as to bias information-seeking in the direction of confirming their hypotheses. Is there a general *tendency* to seek confirmatory information? Or do people prefer *diagnostic* information?

Diagnostic information. The term “diagnostic” is used to refer to information that is useful, relevant, and informative with respect to evaluating the validity of a hypothesis or expectation. A “diagnostic question,” therefore, is one that probes for useful, relevant, and informative responses. Questions and information can vary in their degree of diagnosticity. Consider a situation where a perceiver is testing the hypothesis that a target is athletic. The question, “How many hours each week do you spend exercising?” is more diagnostic than “What did you do after work on Friday?” which, in turn, is more diagnostic than “When did you last get a haircut?” Thus, a preference for diagnostic questions and information is generally viewed as more appropriate and objective; a preference for confirmatory information has often been seen as biased (e.g., Snyder & Swann, 1978).

Confirmation, disconfirmation, and falsification. Seeking confirmatory information, however, is not necessarily biased, irrational, or scientifically inappropriate. There are times when seeking confirmatory information corresponds with seeking the most diagnostic information. Seeking *disconfirmation* sometimes will be less diagnostic than seeking confirmation. This is because there is a difference between (1) seeking for an opposite characteristic than specified by one’s hypothesis (e.g., seeking evidence of introversion, when given an extroverted hypothesis) and (2) seeking information in such a manner as most likely to falsify your hypothesis. I refer to (1) as seeking disconfirmation and (2) as falsification.

This is potentially confusing because (1) both strategies have been considered “falsification” and (2) many researchers accept Popper’s (1959/1968) claim that seeking falsification is one hallmark of the scientific approach to obtaining knowledge. Next, therefore, I try to disentangle all this to provide greater insights into both the interpretation of studies of intuitive information-seeking and the meaning of seeking disconfirmation versus falsification.

Only falsifiable hypotheses can be subject to scientific investigation. “Gravity causes things to fall back to earth” is falsifiable (maybe that apple will just keep going up). “Most college students have high self-esteem” is falsifiable (if most students in a class rate themselves as below average on a list of traits [not likely!], you will have disconfirmed this hypothesis). “UFOs constitute alien visitations” is not falsifiable, because, unless they become IFOs (identified flying objects—but at that point they are no longer UFOs!), it is impossible to obtain evidence demonstrating that there are no aliens up there.

Falsification is important to science for at least two reasons. First, claims that are not falsifiable cannot be evaluated by data. Such claims, therefore, are outside the realm of science (they might be part of morality, faith, or philosophy). Second, the more a theory or hypothesis withstands attempts at falsification, the more confidence we have in the validity of the theory. One could make a pretty strong case that this approach to scientific theory testing provides a good model against which to evaluate the appropriateness of lay, intuitive hypothesis testing (e.g., Nisbett & Ross, 1980; Snyder & Swann, 1978).

Falsification, however, is not the same thing as seeking disconfirmation. This distinction was long overlooked or misunderstood in the lay hypothesis-testing research (see Klayman & Ha, 1987, for a review). It is extremely important, because seeking information in a manner capable of falsifying your hypothesis is not the same as—and is sometimes considerably more informative, objective, and scientific than—seeking disconfirmation in the sense of seeking characteristics that are opposite your hypothesis (Klayman & Ha, 1987).

Consider Mia, who will test the hypothesis that Dennis is *extremely* extroverted. Seeking disconfirmation in the sense of asking questions about extreme introversion is nondiagnostic (uninformative). Showing, for example, that Dennis is *not* extremely introverted would not show that he *is* extremely extroverted. Let’s say Dennis answers “no” to “Do you feel uncomfortable at parties?” This suggests that he is not extremely introverted, but it does not necessarily mean that he is extremely extroverted. Seeking evidence of introversion (disconfirmation) in this sense does not tell us much about whether Dennis is an extreme extrovert.

What type of information-seeking is most capable of falsifying the hypothesis that Dennis is an extreme extrovert? Very few people are wild extroverts. The probability of any one person selected haphazardly or at random actually being a wild extrovert is very low. Probabilistically, therefore, the hypothesis that is most easily disconfirmed is that Dennis is an extreme extrovert.

Questions probing for extreme extroversion, although “confirmatory” in the sense that “yes” responses will confirm the hypothesis, are nonetheless much more diagnostic and scientifically useful than are disconfirming questions. Questions such as, “Are you usually one of the two or three loudest people at a party?” or “Do you usually go to more than three parties per week?” or “Do you frequently style your hair in unusual, attention-drawing ways?” are highly likely to be disconfirmed. Most people will likely answer no to all of these questions. A “yes” response to any of these is diagnostic of unusual extroversion. A “no” response to all three suggests that maybe Dennis is not such an extreme extrovert after all.

Of course, the same probing for extroversion is not as appropriate when testing a more moderate hypothesis. If the hypothesis is that a target is more extroverted than introverted, then testing for extreme extroversion is not that informative. A person could be somewhat extroverted, and yet still give “no” responses to questions probing for extreme extroversion. In this case, a mix of questions probing for introversion and extroversion would appear most appropriate.

Diagnosis and confirmation when testing moderate and extreme hypotheses. Assessing whether intuitive information-seeking conforms to this scientific standard was the issue addressed by another series of studies by Trope and Bassok (1983). In their first study, they asked people to develop yes/no questions probing for either intermediate or extreme politeness and impoliteness. Results showed that people developed highly diagnostic questions. When testing for intermediate politeness or impoliteness, their questions probed for politeness as much as for impoliteness. When testing for extreme politeness, they developed questions that probed for much more evidence of politeness than of impoliteness; when testing for extreme impoliteness, they developed questions that probed for much more evidence of impoliteness than politeness. These are in some sense “confirmatory” questions, but they are also the most diagnostic questions.

They replicated this finding in a second study in which participants were required to choose among prewritten questions that varied in how diagnostic and confirmatory they were with respect to introversion and extroversion. Like the first experiment, they showed that there was no preference for confirmatory questions when testing an intermediate hypothesis, but such a preference emerged when testing for a more extreme hypothesis. Of course, such “confirmatory” questions were also the most diagnostic for testing the more extreme hypothesis. Thus, like Trope et al. (1984), Trope and Bassok (1982) found that people almost always preferred diagnostic questions.

Diagnosing versus confirming strategies in lay information-seeking: Subsequent research. Other studies have compared the diagnostic versus “confirmatory” strategy in a variety of contexts, including personality traits, identifying alien species, and handwriting analysis (Devine, Hirt, & Gehrke, 1990; Skov & Sherman, 1986; Trope & Bassok, 1982). None of these studies found people generating many biased, constraining questions.

In most of these studies, the term “confirmatory” strategy has often been used to refer to questions to which a “yes” would confirm the hypothesis. Even labeling this a “confirmatory” strategy is, in my opinion, misleading. It seems to imply that people constrain others’ responses to confirm their hypotheses when, in fact, these types of questions do not constrain responses at all, and are almost always highly diagnostic.

Regardless, even though “confirmatory” questions may be far less confirmatory than they appear, this subsequent research has consistently found that people overwhelmingly prefer and generate diagnostic questions and information to so-called confirmatory questions and information. Here are the conclusions in the various authors’ own words:

Trope and Bassok (1982, pp. 30–31): “[Our] three studies provide strong evidence that people gather information by what we termed the diagnosing strategy . . . [and] . . . very little evidence for interest in questions about features whose presence tends to confirm rather disconfirm the hypothesis, as the confirming strategy postulates” (emphasis in original).

Skov and Sherman (1986, p. 111): "When given a choice among questions that differed in diagnostic value, subjects almost always chose those that were more diagnostic. In fact, diagnosticity was the main determinant of question selection."

Devine et al (1990, p. 960): "Diagnosticity clearly plays the primary role in trait hypothesis testing. Subjects clearly showed an overwhelming preference for highly diagnostic information in this research."

Nonetheless, across all of these studies (Devine et al., 1990; Skov & Sherman, 1986; Trope & Bassok, 1982), there was also a tendency for people to prefer questions to which a "yes" response would confirm the hypothesis over questions to which a "yes" response would disconfirm the hypothesis. The extent of this preference was most clearly demonstrated in the third study of Devine et al. (1990), in which questions were equated for diagnosticity. On average, if there was no preference for so-called "confirmatory" questions, people should generate hypothesis-true questions (i.e., questions to which a "yes" would confirm the hypothesis) as often as alternative-true questions (questions to which a "yes" would disconfirm the hypothesis). That is, there should be a 50–50 split between hypothesis-true and alternative-true questions.

Devine et al. (1990) did not find a 50–50 split. Fifty-six percent of the questions people preferred were hypothesis-true questions. People asked lots of questions, so this departure from 50% was whoppingly statistically significant. But is it a whopping preference for "confirmatory" questions? I guess that's a matter of subjective opinion. But what we are talking about here is a grand total of a 6% departure from completely even-handed questioning. There does seem to be some tendency to prefer "confirmatory" questions, but that tendency looks pretty small to me.

Is There Any Bias in Social Information Gathering?

The bias hypothesis has a near-death experience. The research reviewed thus far might convey the impression that Snyder and Swann (1978) were almost completely wrong, and that their evidence for confirmatory social hypothesis testing was an idiosyncratic finding that resulted from unnatural and artificial aspects of their procedures. People apparently have a very strong preference for asking diagnostic questions. When asking diagnostic questions, they do have a slight tendency to prefer questions to which a "yes" will confirm their hypothesis (over questions to which a "no" will confirm their hypothesis), but even this preference is quite small. So, you may be wondering, "Is there any bias at all in lay intuitive social hypothesis testing and information-seeking?"

To evaluate this question, we must first figure out which evidence is relevant. Every study that has allowed people to generate their own questions has shown that people almost never create questions that constrain targets' answers to confirm the hypothesis (Swann & Giuliano, 1987, Experiment 1; Trope et al., 1984). If people do not create such questions, then the research that has required participants to select from lists of patently biased and constraining questions (e.g., Fazio, Effrein, & Falender, 1981; Snyder & Swann, 1978; Swann & Giuliano, 1987, Experiments 2 and 3) probably does not tell us much about what goes on in most naturally occurring interactions.

Zuckerman, Knee, Hodgins, and Miyake (1995) breathe new life into a heavily diluted bias hypothesis. All this would seem to suggest that Snyder and Swann's (1978) early study was a red herring, because people go about gathering social information in a manner far less biased and more appropriate than their study seemed to suggest. But such a conclusion would be too strong. It turns out that, perhaps in dramatically diluted form, Snyder and Swann's (1978) original conclusions and hypotheses can be revived.

This is mainly because of something known in technical/methodological circles as "acquiescence." Researchers studying how people respond to questionnaires have long known that sometimes some people tend to respond to almost any question with a "yes" or "agree" response (e.g., Lenski & Leggett, 1960; Schuman & Presser, 1981). Similarly, in many interpersonal situations, it is often easier to agree than to disagree. Disagreeing is, well, disagreeable. And especially if the questions involve fairly vague or ambiguous issues or features, it is often not too hard for most people to come up with sufficient justification to agree with others (Krosnick, 1991).

I can almost hear the wheels turning in your head. "Aha!" you may be thinking, "If people tend to ask questions to which a 'yes' answer confirms the hypothesis, and if there is some general tendency to respond 'yes' regardless of question, then information-seeking may still be biased in the direction of confirming the perceiver's hypothesis." Even if you were not thinking that, Zuckerman et al. (1995) did, and performed a study to examine whether this flow of events actually occurred.

Zuckerman et al. (1995) used the same basic question-creation/interview methodology used in the prior studies. Interviewers asked interviewees four questions about each of four traits (16 questions total): trust, calm, extroversion, and optimism. Some were asked to probe for the negative version of the trait (e.g., suspicious), whereas others were asked to probe for the positive version of the trait (e.g., trusting; a third group was given a double hypothesis to test, e.g., to find out whether the person was suspicious or trusting, and their result tended to fall between the other two, although there was a tendency to be close to the group testing the more positive hypothesis).

Did people tend to ask questions to which a "yes" would confirm the hypothesis? (Although prior research had referred to this as a "confirmatory strategy," following Klayman and Ha's [1987] extremely lucid analysis of information-seeking, Zuckerman et al. [1995] referred to this as a "positive test strategy," and I will do so for the remainder of my discussion of their study.) People did use a positive test strategy. On average, when testing the positive end of the trait, interviewers asked one more question to which a "yes" would confirm the positive trait than did people testing the negative end of the trait. Concretely, this means that, for example, people asked to test for extroversion might be more likely to ask a question such as, "Do you make friends easily?" than would people asked to test for introversion.

Was there acquiescence? Indeed, there was. Across all conditions, 59% of all responses were "yes." Even when interviewees were responding to interviewers probing for negative traits (who were slightly more likely to ask questions to which a "yes" would confirm presence of that negative trait), 54% of responses were "yes." Sixty-one percent of responses to interviewers probing for positive traits were "yes."¹

Did acquiescence combine with the positive test strategy to lead to interview data that was biased in the direction of confirming the hypothesis? It did. Interviewers probing for the positive end of the trait (e.g., trusting, calm, extroverted, or optimistic) were more likely to

ask “positive trait yes” questions; interviewees (regardless of interview question) were more likely to say “yes” than “no”; and, as a result, interviewers probing for positive traits were more likely to receive evidence of that trait than were interviewers probing for negative traits. On average, about three of every four responses received by interviewers testing for a positive trait confirmed that the interviewee had the trait, whereas only about two and a half of every four responses received by interviewers testing for a negative trait indicated that the interviewee had the positive trait. Thus, interviewers testing for the positive side of the traits evoked more positive responses from the interviewees than did interviewers testing for the negative side of the same traits.

This is a self-fulfilling prophecy (a false positive expectation evoked more positive responses). Thus, after all, it looks as if expectancies can bias question asking, and biased question asking can lead targets to provide answers that disproportionately support the original expectation. How powerful was this effect? The correlation between perceiver hypothesis and positivity of target response was .29—right in the 0.1 to 0.3 range typical of self-fulfilling prophecies (Jussim, 1991; Rosenthal & Rubin, 1978).

The bias hypothesis: Revived but weaker than it may seem. Zuckerman et al. (1995) shed a great deal of light on some of the complex and controversial issues surrounding lay hypothesis testing. Nonetheless, their 0.29 effect may overstate the self-fulfilling effects of lay social information-seeking. Here is why.

Swann and Ely (1984), in the “Battle of Wills,” self-verification versus self-fulfilling prophecy study described in Chapter 7, also examined the relationship between the types of questions perceivers asked regarding introversion and extroversion and judges’ ratings of targets’ intro-/extroversion. That correlation was nearly zero (0.09, which was not significantly different than zero). In other words, with respect to overt behaviors detectable by independent judges, perceivers’ questions had little or no self-fulfilling effects at all.

Why this difference between the results of Zuckerman et al. and Swann and Ely? Although pending further research the answer to this question must be speculative, there is at least one strong contender—they had different outcome variables. In the Swann and Ely (1984) study, judges’ overall impression of targets’ intro- or extroversion constituted the evidence of self-fulfilling prophecy. In contrast, in the Zuckerman et al. (1995) study, judges never provided such an overall impression. Instead, Zuckerman et al. (1) required perceivers to ask questions that could be answered with “yes” or “no” responses and (2) tallied the number of “yes” and “no” responses (which constituted the evidence of self-fulfilling prophecy).

Thus, whether the small tendency of perceivers to evoke expectancy-confirming responses from targets, as found by Zuckerman et al. (1995), is sufficient to lead anyone (perceivers, targets themselves, or outside judges) to view targets as confirming the expectation is unclear. Just because Louise is slightly more likely than Louis to say she enjoys big parties, it does not necessarily mean that Louise will be seen as much more extroverted than Louis (especially if the difference is small enough).

In addition, whether the small tendency to provide more “yes” responses to positive test questions constitutes much of a self-fulfilling prophecy is unclear. It is, at most, a demonstration of a very superficial self-fulfilling prophecy, because it is entirely based on verbal responses to a single interviewer’s questions. Whether such verbal responses have any enduring effects is unclear. Considering the targets of an extroverted hypothesis, we do not know, for example, whether they would provide similarly extroverted verbal responses to a new,

no-expectancy interviewer, or whether they would act in a more extroverted manner with other people.

Regardless of whether one takes the Zuckerman et al. (1995) study at face value, or tempers its conclusions with Swann and Ely's (1984) finding of a very weak relationship between question asking and impressions, or further tempers their conclusions with a consideration of the unknown endurance or generality of the self-fulfilling prophecy effects Zuckerman et al. (1995) did find, the overall pattern is consistent with one of the major themes of this book. Self-fulfilling prophecies and expectancy-based biases are real, but they are typically small, fleeting, and weak.

Conclusion: The Less Than Awesome Power of Expectations to Distort Information-Seeking

Do people's expectations bias the manner in which they seek social information? Do people rig the interaction, intentionally or not, to get what they expect? The answer is a resoundingly clear "Ahh, well, uh, kinda sorta maybe a little."

The following conclusions are justified by the research on the role of expectations in social information-seeking: (1) People almost never spontaneously ask the type of biased, constraining questions that Snyder and Swann (1978) required perceivers to use in the first study of lay social hypothesis testing; (2) in general, people greatly prefer diagnostic questions and information over confirmatory information and questions; (3) there is a slight tendency for people to prefer questions to which a "yes" answer confirms the hypothesis over questions to which a "no" answer confirms the hypothesis; (4) from a scientific or logical standpoint, such questions are often, though not always, highly diagnostic and appropriate; and (5) the combination of perceivers' use of a positive test strategy with targets' tendency to give more "yes" than "no" answers (acquiescence) may ultimately lead perceivers to obtain social information in a manner somewhat more likely to confirm than disconfirm their hypotheses.

Bottom line: Expectancy effects, both self-fulfilling prophecy and bias, are real, but people are not completely out to lunch. In fact, they are hardly out to lunch at all. Mostly, they are minding the store quite effectively. They do not seek to blindly confirm their prior beliefs. Although biases do creep in, people prefer accurate information and do not rigidly resist disconfirming information. All of this may help explain why expectations have some, but typically not all that much, power over perceptions, judgments, and evaluations—as will be discussed in the next chapter.

Note

1. Overall, 59% of all responses were "yes," even though 54% of responses to all probes for negative traits were "yes" and 61% of all responses to probes for positive traits were "yes."

This was because there were more probes for positive than for negative traits.

9 The Less Than Awesome Power of Expectations to Bias Perception, Memory, and Judgment

CHAPTERS 6, 7, AND 8 SUGGESTED that self-fulfilling prophecies were not quite as powerful as the early research and many of the early reviews seemed to suggest. This chapter complements those chapters by considering the extent to which the early research demonstrated that expectations lead to biases in the mind of the perceiver. I use the term “bias” to refer to an influence of perceivers’ expectations on *their own* subjective judgments, *not* on objective reality (creating an objective social reality—changing *targets* in expectancy-confirming ways—is self-fulfilling prophecy). Nonetheless, this chapter is similar to the prior ones in that I suggest that, just as the early research did not justify conclusions emphasizing the power of expectations to create objective social reality, the early research did not justify conclusions emphasizing the power of expectations to bias perceptions, evaluations, and memory.

This chapter follows a format much like that of Chapter 6, on self-fulfilling prophecies. It has two major sections that take a closer and more critical look at the potentially biasing power of expectations in two very different ways. In the first section, I present a series of common experiences in daily life in which biases either do not occur, occur to only a modest extent, or occur only infrequently. This is important both to help build at least a *prima facie* case against ascribing any sort of inevitability or great power to expectancy-confirming biases and to link my research-based perspective on the limited power of expectancy effects to frequent everyday events. The second section revisits some of the most highly cited and classic expectancy bias studies from this early period (again, roughly 1970 through 1990) in order to evaluate just what they do and do not say about the power of expectancy effects.

On the (Non-)Inevitability of Expectancy-Confirming Biases: Some Examples From Everyday Experience

COACHING SOCCER

When my daughter, Rachel, was in second grade, I began my career as a soccer coach. New coaches tend to get the newer (and younger) kids, not (as far as I can tell) because of elbow-rubbing old boy bias favoring existing coaches, but because the leagues usually try to keep many of the kids together. By definition, therefore, more experienced coaches tend to keep more of the more experienced kids. So, I had a young, small team.

Of course, they lost the first game. They also lost the second. And the third. In fact, we did not score a single goal until the next to last game of the season (we lost that game 5–1). We did pick up a 0–0 tie somewhere in the middle of the season. So, by the time the regular season was over, we had no wins, eight losses, and one tie.

But then came the playoffs, which every team automatically entered. Playoff games differed from regular games, however, in that they could not end in a tie. Should the score still be tied at the end of a full game, the winner is determined by a “shoot-out.” In regular shoot-outs, for each team, five players (one at a time) shoot on the opposing goalie. At the end of the 10 shots, whichever team has the most goals wins.

The procedure was basically the same for our playoffs, with one difference. Should there be a shoot-out, it would be done with no goalie (these were first- and second grade kids).

Knowing that we had a shot at a game ending in a 0–0 tie, in preparing for these playoffs, we decided to have our kids practice the shoot-out. While they were practicing, the other coach and I were discussing which five kids we would select (there were nine kids on the team). Our initial inclination was to go with the biggest, strongest kids. They had more of a soccer look and were generally the best players on the team. Had we done this, this would have been a classic expectancy effect (size-based athletic competency expectancies coloring our judgments of who was most likely to do well in a shoot-out).

But then we decided to have a competitive dress rehearsal in practice. After letting them all practice shooting for about 10 minutes, we gave each kid nine shots (three at a time). The five kids with the most goals would be the ones chosen for a shoot-out. As it turned out, one of the big kids had a very strong kick, but could not control it very well, so she did not score very many goals in this practice shoot-out. Then there was this other small kid, who had very weak, but very accurate, kicks. She would be highly unlikely to score in the regular part of a game, but with no goalie, she ended up ranking fourth in the practice shoot-out.

So, what should we do? Should we keep the weak-hitting small kid in and hope that her success in the practice was not just luck? After all, the bigger kid who did not make it was an excellent player and came from a very athletic family (e.g., her dad was an assistant coach on the Rutgers football team at the time). The classic social psychological perspective on perceptual expectancy effects (see Chapter 5) would seem to predict that (1) we would remember the bigger kid's kicks as better than they really were (e.g., “they missed, but they were awfully close”—see, e.g., the Darley & Gross [1983] study described in Chapter 5); (2) we would make expectancy-confirming attributions (“the big kid had bad luck; the little kid just got lucky”—see, e.g., the Kulik [1983] or Deaux & Emswiller [1974] described in Chapter 5); or (3) we would possibly even reconstruct our memory of what happened to misremember the

bigger kid as having made more goals than she really did or to misremember the smaller kid as having made fewer goals than she really did (see, e.g., the Snyder & Uranowitz [1978] study described in Chapter 5).

In fact, however, none of this happened. When the competitive practice was over, I looked at the assistant coach; he looked at me. I said something like, “Well, that was useful, both for them and for us.” Now, you might be thinking, “Well, of course, you [referring to me] are a social psychologist; you are familiar with people’s tendency to allow their expectations to bias their judgments. It is no big deal if you were alert enough to ward off such effects. This is, therefore, not a great example of little or no expectancy-maintaining bias effects.”

This, however, would not explain my assistant coach’s reaction. He looked back at me and said, “Amanda [the smaller kid] is in; Dori [the larger kid] will play defense.” Not the slightest hesitation. Not the slightest bit of resistance to my initially gentle suggestion that competition upended, in part, our expectations. (Actually, while we are on the topic of accuracy, the other four kids who scored the highest in the practice shoot-out were all kids we expected to make it. So the predictive accuracy of our expectations was actually very high, although not perfect—but accuracy is a topic I will leave for subsequent chapters.) I did not have to convince him or persuade him in any manner. He had reached the identical conclusion. Amanda was in; Dori was out (although Dori’s kicks were not all that accurate, she was the bulwark of our defense, in part because her kicks were so strong—even though we lost all those games, most were only by scores of 1–0 or 2–0). Not much in the way of expectancy-maintaining bias here.¹

And boy did this pay off. After losing to the top team in the league (the playoffs were double elimination), we managed to eke out a 2–1 victory over a middlin’ team. Then we faced another strong team. A grueling, gripping, defensive battle ended in a 0–0 tie. This was exactly what we had prepared for.

Thus commenced the shoot-out. Their first kid scored; our first kid scored. Their second kid scored; our second kid missed. We were down 2–1. Their next kid missed; our next kid scored: 2–2. Their fourth shooter was the opposing coach’s daughter and one of the best players in the league. She shot a hard kick that missed. Up to the shooting line went little Amanda. She ran and kicked. The ball slowly dribbled toward the goal. It gently hit the inside of the goal post. It rolled parallel to the goal for about a yard, so slowly you could make out the writing on the ball. It then hit a clump of grass, which nudged the ball, barely, over the goal line. She scored. Both teams’ next shooters missed, so we won 3–2, in part precisely because expectancy-maintaining bias *did not* have much ultimate influence on our evaluation of our shoot-out players.

OTHER EXAMPLES

Daily life is filled with examples of people changing their expectations to fit the evidence, rather than changing their memory, evaluation, or attribution for the evidence to maintain the expectation. Before the 2000 World Series, tons of Mets fans would call in to local talk radio stations claiming that the Mets were better than the Yankees and would win the series. After the series (which the Yanks won), there were no more such calls.

Research faculty are often faced with students who surprise them—in either the positive or negative direction. Many of us have had experience with graduate students with great

undergraduate records, including high GREs, high GPAs, tons of research experience, etc., who suffer through a few years of graduate school and then drop out. Although we faculty may still evaluate their *potential* highly, when a master's thesis is still not completed after more than 3 years, when there are neither publications nor conference presentations in the same time period, and when most graduate course grades are in the B range, few of us still think of them as star students. Similarly, most of us are fully capable of recognizing achievement as unusually strong when a student knocks off a master's thesis in less than 2 years, quickly completes other research projects and submits them for publication, etc.—even if we did not have him or her pegged as a star student when he or she first arrived.

Chapter 6 discussed the non-self-fulfilling nature of the stock market most of the time. Revision of expectations in response to evidence happens all the time in the stock market. Consider companies expected to grow very quickly. Such companies usually have high and rapidly growing stock prices. However, in general, not long after growth slows, the stock price takes a nosedive. Why? Because one slow-growth quarter means that investors cannot be sure of continued high growth. A declining stock price in this context, therefore, reflects lowered investor expectations.

I am not denying that expectancy-maintaining biases occur. Coaches probably give more of a benefit of a doubt to a high-expectancy player who puts in a bad performance or two than to a low-expectancy player who plays equally poorly. Same thing for faculty and their advisees. Sports fans probably are more likely to see close or ambiguous umpire/referee/official calls as favoring their preferred teams. And there are usually investors buying stocks as their prices decline, at least in part because some may not have revised their expectations even in response to the quarter's slow growth.

Thus, expectancy-confirming biases do happen in real life. Sometimes, such effects are quite large. In general, however, they tend to be small and fleeting and occur primarily (although not necessarily exclusively) in situations in which social reality is unknowable, unclear, or ambiguous. Even when social reality is unknowable, unclear, or ambiguous, however, expectancy-confirming biases are neither inevitable nor, on average, particularly powerful. Furthermore, daily life is filled with examples of people's expectations, rather than dramatically biasing their perceptions of reality, flexibly changing in response to reality. It happens in the classroom, on the athletic field, in the stock market, and pretty much in every context I am aware of.

Of course, this set of claims regarding the limited nature of expectancy-confirming biases is not restricted to either common sense or my personal evaluations of everyday experiences. Indeed, it is not even primarily based on such considerations. Instead, it is based primarily on a critical analysis of what can and cannot be concluded on the basis of the empirical research on expectancies. The next section, therefore, revisits many of the early classic studies of expectancy-confirming bias and demonstrates how *even those studies* typically demonstrate only weak expectancy-confirming biases.

The Classic Stereotype-Based Expectancy Bias Studies Revisited

As in previous chapters, this review is selective rather than comprehensive, because there are literally hundreds of studies that have examined the ways in which expectations and

stereotypes bias judgments, perceptions, evaluations, and memories for specific targets (see, e.g., reviews and meta-analyses by Darley & Fazio, 1980; Fiske & Neuberg, 1990; Jussim, 1991; Kunda & Thagard, 1996; Stangor & McMillan, 1992; Swim, Borgida, Maruyama, & Myers, 1989). Those included here once were, and often still are, highly cited in major reviews of expectancy effects. Many often pop up in textbooks and reviews as examples of how stereotypes and prejudice can lead to discrimination and bias.

Therefore, the articles reviewed in this chapter are only a small, select, and biased sampling of research on expectancy-confirming bias from the early 1970–1990 period—but that bias is entirely in the direction of including studies commonly interpreted as emphasizing the power and pervasiveness of expectancy-confirming biases. Indeed, I have purposely tried to focus on the most influential expectancy-confirming bias studies from this early era.

Why focus on such studies? Because, as far as I can tell, *even such studies* fail to justify an emphasis on the power and pervasiveness of expectancy effects. Daily life provide numerous examples of limited or nonexistent expectancy-induced biases, and even studies frequently cited in testaments to the power of expectations actually provide little such evidence.

There are, however, a small number of exceptions to my 1970–1990 restriction. First, in a few rare cases, in later years, researchers have performed nearly exact replications of classic studies from this earlier period or studies that, although methodologically different, addressed the exact same issue. When that is the case, in order to evaluate how well the conclusions from the early classics have stood up over time, I also review the later research. In addition, I also discuss the results of several meta-analyses published after 1990. Although those meta-analyses were published after what I have termed the early years of social psychology's love affair with expectancy effects, they drew primarily on research published during that time—thus, they provide an excellent panoramic snapshot of the conclusions that are justified by the research from that early period.² By helping to answer the question, “How well have the conclusions reached on the basis of a small number of dramatic and highly influential early studies stood up over a long haul that has included testing the bias hypothesis hundreds of times in a vast variety of ways?” both the replications and the meta-analyses help fill in a major piece of the expectancy puzzle. That said, let's revisit the early bias classics.

RACIAL STEREOTYPES

The saga of Duncan (1976). This was the African American/White ambiguous shove study described in Chapter 5. The main result was that when the African American student shoved the White student, 75% of the viewers described the action as “violent”; when the White student shoved the African American student, only 17% described the action as violent. This huge difference is completely consistent with many social scientists' suspicions about the nature of modern White prejudice—few people go around saying, “I hate Blacks” or that discrimination and segregation should be legalized. Instead, their prejudice “leaks” out when they are not thinking about it. One manifestation of such leakage is interpreting ambiguous behavior in a stereotype-consistent manner—especially ambiguously hostile or aggressive behavior.

Seventy-five percent versus 17%? That is huge. It is so huge that I am not aware of any other study finding such a large effect of stereotypes or prejudice on person perception. None of

the other studies described in Chapter 5 produced such a huge effect. Even other studies similarly testing for hostility-based racial prejudice have not found anything remotely resembling such huge effects (e.g., Devine, 1989; Prentice-Dunn & Rogers, 1980; Sagar & Schofield, 1980).

In fact, few social psychological experiments of any kind produce such large effects. The effect size associated with that difference would approximately equal a correlation of .6 between race/perpetrator combination and judgments. It is especially unusual because, as far as I can tell, this is the only article Birt Duncan ever published. This is relevant because it would be far less surprising to me for a well-established researcher working on a line of research for years to have so honed in on a phenomenon and so refined a set of procedures to have crafted an experiment that produced such a large effect. But a relative novice? Especially a relative novice who never published another article? Not impossible, but certainly very unusual.

These thoughts long led me to suspect something was fishy about this study. I first came across this study when I was in graduate school in the 1980s. It looked odd to me, but that seemed to be the end of it. Then I moved to Rutgers in 1987, where I worked happily for years. In 1999, I brought this study up in casual conversation with a colleague, Professor Richard Ashmore, whose office was just a few doors down from mine (he has since retired). He had briefly worked with Birt Duncan and told me the following story.

Duncan was in the social psychology program at Princeton in the early 1970s. He had been an undergraduate at UCLA, where he worked with Barry Collins, Ashmore's graduate advisor. With this in common, they met and, with Professor Mel Gary (also at Rutgers but now retired), decided to work together on a study of stereotypes. They ran a study that was nearly identical to the one described in Duncan (1976), with the following major differences: (1) It was conducted in New Jersey, rather than California, and (2) the study did produce a result much like Duncan reported, but with a much smaller effect.

One of the main results reported in Duncan (1976) is that African American perpetrators were seen as more aggressive than White perpetrators regardless of the race of the victim. When Ashmore, Duncan, and Gary³ found this, they considered two possible explanations. The first was their preferred one—that stereotypes biased perceptions of the African American perpetrators. The second, however, was that the African American actors just did a more credible job of appearing angry or hostile. In reviewing the tapes, they concluded that the second explanation seemed more likely and, at minimum, could not be ruled out. Therefore, they did not publish the study.

Years later, Duncan had apparently moved to UC-Berkeley and published a study that was nearly identical to the one run with Ashmore and Gary, complete with a footnote thanking their contribution to “pilot work,” and which produced the picture-perfect results reported in the 1976 article. As far as I can tell, Duncan then disappeared from social psychology. Whether the data can be believed, however, goes beyond the uncanny methodological similarity with the study performed with Ashmore and Gary, and beyond the striking perfection of the results.

Even taking the study at face value, a close examination of the published report yields two additional sources of concern. First, nowhere in the method section does Duncan state that the confederates were rigorously trained in order to equalize and standardize their behaviors during the interaction leading to the shove. This is a highly unusual omission—research

using confederates almost always includes such training, which is generally mentioned explicitly in the method section (e.g., Chen & Bargh, 1997; Jussim, Soffin, Brown, Ley, & Kohlhepp, 1992; Word, Zanna, & Cooper, 1974). This omission is, of course, ambiguous. It may mean that the confederates' behavior was not standardized, but perhaps it was an innocent omission.

Second, on page 594, there is a section titled "Quality of the Stimulus Tapes" that reports that Duncan (1976) had 40 high school students serve "as treatment blind judges" whose job was to "rate" each person in the tape on characteristics like moral, aggressive, hostile, etc. Duncan (1976, p. 594) reported: "Analysis of the treatment blind judges' ratings of the black and white confederate for each stimulus tape revealed no between-condition differences."

This section was included because Duncan (1976) needed to argue that there were no real objective differences between the behaviors of the actors in the tapes. Any perceived differences, therefore, could only represent differences occurring in the minds of the perceivers, rather than in the actual behavior of the targets. But it is not clear to me what this actually means.

How did he establish an absence of real differences? With judges described as "treatment blind." So, how were these judges rendered "treatment blind"? (The phrase "treatment blind" means that judges were not aware of which experimental condition was represented in each tape.) If they are viewing the tape, why couldn't they see the race of the interactants for themselves? Duncan (1976) gave no explanation. So, I am left wondering, why did the college "subjects" see huge Black/White differences, but the high school "judges" did not? If the high school judges were aware of the interactants' race, this would appear to be a replication that failed miserably. And how could they have not been aware of their race?

I am not claiming that Whites are never biased against African Americans or that stereotypes never influence person perception. However, given reasonable doubts about the replicability of Duncan (1976) based on the implausibly high level of perfection of his results; on the weaker pattern of results obtained when Ashmore, Duncan, and Gary ran a highly similar study; and on the sketchy results Duncan (1976) himself reported regarding high school students, this study probably provides less than ideal support for claims about the power of expectations to influence social perception. Indeed, in my view, the doubts about the replicability of the study are sufficiently severe as to warrant a moratorium on citing it pending replication of the large differences he obtained.

Subsequent research. Of course, even if Duncan's (1976) results can be taken at face value, it is important to remember that it is only a single study. Thus, you might be wondering, "Well, what have other studies of the role of racial stereotypes in person perception found?" By 1990, dozens of studies had examined exactly this question. And it is clear that Duncan's (1976) pattern of large bias has not held up.

When targets' behavior is highly ambiguous or when perceivers can seemingly justify biased responses with something other than prejudice, small biases have sometimes emerged. For example, Sagar and Schofield (1980) ran a replication of Duncan's (1976) study among sixth graders. They found a similar pattern, in that ambiguously aggressive actions (bumping in the hallway, requesting food, etc.) were seen (by both African American and White students) as more aggressive when committed by an African American student than when committed by a White student.

But the bias effect size (correlation between race of actor and aggressiveness ratings) was a mere 0.23. In intuitive terms, this means stereotypes changed the interpretation of the targets' actions about 12% of the time.⁴ Twelve percent ain't nothin'. It clearly represents bias. Such bias could be socially important. But it also means that stereotypes *did not change* people's perceptions about 88% of the time.

Furthermore, when targets' behaviors or accomplishments are clearly positive or negative, or when objective relevant information is abundantly available, little or no evidence of anti-African American bias emerges (e.g., Feldman, 1972; Jussim, Coleman, & Lerch, 1987; Jussim, Eccles, & Madon, 1996; Linville & Jones, 1980; Madon et al., 1998; McKirnan, Smith, & Hamayan, 1983). Indeed, in these studies, there was, on average, a tendency to *favor* African Americans over Whites with similar characteristics. In general, it had long been known within social psychology that people's attitudes toward people from other racial groups were typically determined far more by *belief similarity* than by race (Cook, 1984; Rokeach & Mezei, 1966). Thus, (1) although bias against African Americans may be alive and kicking (see, e.g., Devine, 1989; Jones, 1996), such bias is far from inevitable; (2) even when it appears, it rarely, if ever, packs the type of punch reported in Duncan (1976); and (3) referring exclusively to studies of the role of racial stereotypes in person perception, reverse biases favoring African Americans appear in some studies, whereas biases favoring Whites appear in others.

SOCIAL CLASS STEREOTYPES

The saga of Darley and Gross (1983) and Baron, Albright, and Malloy (1995). Darley and Gross (1983) was the study of social class stereotypes described in Chapter 5, in which Princeton students evaluated a fourth grade girl's performance and ability much more positively when they believed she was from a middle class (rather than lower class) background, if they observed her test performance (little or no bias without the test performance).

This was a well-designed study, with none of the ambiguities surrounding Duncan's (1976) research. Although I do not doubt that Darley and Gross (1983) found what they found, it has proven difficult to replicate both the finding that social class biasing person perception in the presence of relevant individuating information and the pattern of no bias without individuating information.

The only research (Baron et al., 1995) to attempt exact replication of Darley and Gross (1983) failed. Baron et al. (1995) actually performed two replications and extensions of Darley and Gross (1983). The first one was identical to the original study (they even obtained the same tapes for manipulating social class from Darley), except for the following differences. Perceivers were not Princeton students. Instead, they were New England college students attending either a large state university (in Study 1) or an elite private college (Study 2). In addition to the no-performance condition, Baron et al. created three versions of the tape of the fourth grade girl taking a test—low, medium, and high performance (getting 25%, 50%, and 75% of the questions correct, respectively, on the Wechsler Intelligence Scale for Children). This was to directly test Darley and Gross's (1983) suggestion that stereotypes are most likely to bias perception when individuating information is ambiguous. If so, they reasoned, then perhaps there would be bias in the 50% performance condition, but not in the others.

Their first set of analyses focused exclusively on the no-performance versus 50% correct conditions—the set of conditions corresponding almost exactly to that used by Darley and Gross.⁵ What did they find? Baron et al. did not merely fail to replicate Darley and Gross's pattern—they found *the exact opposite*. Specifically, Baron et al. found that perceivers *did* evaluate the fourth grade girl's intelligence and ability more positively when they believed she came from an upper class background, *but only in the absence of performance information*. The girl's supposed social class had no influence on perceivers' judgments of her intelligence and ability in the presence of performance information.

What about the other conditions? Baron et al. performed a second set of analyses examining whether social class biased judgments across the three performance conditions. It did not. The girl seen playing in the poor urban playground was rated as just as smart as the (same) girl seen playing in the middle class, suburban playground (for the statistically inclined, this means that there was no main effect for social class in this analysis, nor was there an interaction between social class and performance).

Instead, judgments were solely influenced by the girl's performance. In other words, perceivers rated her as most intelligent when she got 75% correct, as middling when she got 50% correct, and as least intelligent when she got 25% correct. Apparently, people were not quite as hell bent on confirming their expectations as the research discussed in Chapter 5 seemed to suggest. Indeed, they seemed downright responsive to objective social reality. Just to be sure, Baron et al. (1995) repeated their study at a small, highly selective, and overwhelmingly middle- and upper class private college in New England. The results from their own Study 1, not from Darley and Gross (1983), replicated almost exactly.

Of course, it is possible that Darley and Gross's (1983) claim about how people use stereotypes is basically correct *and* for Baron et al. (1995) to have failed to replicate their findings. How? Perhaps the students Baron et al. (1995) studied did not hold quite as powerful (extreme, confidently held, etc.) social class stereotypes as did the Princeton students Darley and Gross (1983) studied. As an Ivy League school largely for the economic elite, the plausibility of there being unusually strong social class stereotypes at Princeton seems quite high. Thus, perhaps weak stereotypes have little or no effect on person perception (as per Baron et al.) and stronger ones work as Darley and Gross suggested.

Regardless of whether this latter analysis is true, however, the mere fact that Baron et al. (1995) failed to replicate Darley and Gross (1983) raises the possibility of either of two major limitations to Darley and Gross's conclusions: (1) Perhaps Darley and Gross's (1983) findings were a fluke coincidence, never to be replicated again; (2) perhaps Darley and Gross studied social class stereotypes among a group (Princeton students) where such stereotypes happen to be particularly strong, but most people do not rely on social class stereotypes in judging others as much as did their Princeton students. In either case, reaching a general conclusion about the power of stereotypes based on Darley and Gross (1983) appears largely unjustified.

Other research. Of course, perhaps the results from Baron et al. (1995) are the fluke coincidence, and future research will ultimately confirm Darley and Gross's (1983) perspective. There are reasons, however, to believe that such an outcome is not likely. First, in contrast to Darley and Gross's (1983) finding that people did not judge the student based on her social class when they had *only* social class information, numerous experiments show that people *do* judge others based on their social class, in the absence of much other information about their personal attributes or accomplishments (e.g., Bayton, McAllister, & Hamer, 1956; Coleman

et al., 1995; Feldman, 1972; Jussim, Coleman, & Lerch, 1987; Jussim, Fleming, et al., 1996; Smedley & Bayton, 1978). In general, without much more to go on other than social class, people have higher expectations for middle- and upper middle class people than for working-class and poor people.⁶

One source of the appeal and drama of the Darley and Gross (1983) study, at least in the retelling, is its apparent relevance to real-world social problems associated with social class. For example, if the cognitive processes identified by Darley and Gross (1983) are widespread and general, then one would expect teachers to show similar biases in their evaluations of students. But they don't. Naturalistic research consistently shows that teachers' judgments of students (e.g., the grades they assign) are not biased against students from lower social class backgrounds (Jussim et al., 1996; Madon et al., 1998; Williams, 1976). These are the only studies of which I am aware that have quantitatively and objectively studied the role of social class stereotypes in person perception in a real social context of any importance (education). And they all essentially replicated the Baron et al. (1995) results showing no bias in the presence of individuating information, rather than the Darley and Gross (1983) results showing bias in the presence of individuating information.

SEX STEREOTYPES

The saga of Goldberg (1968). This was the famous "women evaluate male authors more favorably than they evaluate female authors, even when they write the exact same thing" article discussed in Chapter 5. This article was a favorite of the early expectancy perspectives, because it so simply and clearly demonstrated a biasing effect of sex stereotypes. The fact that Goldberg (1968) studied only women, I suspect, only increased the impact of the study. Rather than viewing it as a limitation ("maybe the results would not apply to men"), I suspect that the general reaction was more like "imagine, if women are this biased, men are probably even more biased!"

But how this study was once interpreted no longer really matters. Why? Because this result has not held up, either. By the early 1990s, dozens of studies had examined whether sex stereotypes bias people's judgments of men's and women's work or accomplishments.

Swim et al. (1989) took advantage of this huge database and performed a meta-analysis. So what was the average difference in judgments of men's versus women's work in the 119 studies Swim et al. (1989) examined? It was nearly zero. The correlation between target sex and evaluations was $-.04$ (the negative sign indicates a tendency to favor men). Because of the large number of studies, this $-.04$ was statistically significant, meaning that there was indeed a consistent tendency to favor men. But that tendency, though statistically greater than zero, was very small. It is the equivalent of people favoring men once out of 50 comparisons between men and women (Rosenthal, 1985). This, of course, is the same as saying there is no sexism in 49 of 50 comparisons.

Furthermore, Swim et al. (1989) divided the 119 studies up in dozens of ways in an effort to find at least one condition under which bias was large (e.g., male vs. female subjects, target sex, sex-role orientation, attractiveness or competence). They failed. No analysis that included at least five studies yielded an effect size greater than 0.2.⁷ Goldberg (1968) was not the only researcher to find bias against women; however, for every study like Goldberg's, there were almost as many finding bias against men.

Sex stereotypes do sometimes bias judgments: .04 is not zero. Politically, the argument that any bias is unjustified is pretty persuasive. And even though Swim et al. (1989) failed to find any conditions under which such bias was powerful, I do not doubt that, occasionally, such biases can be powerful. But the Swim et al. (1989) meta-analysis clearly showed that such effects are not generally very powerful.

THE MENTAL ILLNESS LABEL

The saga of Rosenhan (1973). This was the pseudopatient study (described in Chapter 5) purporting to show that the sane were not distinguishable from the insane, and which has long been cited as a classic example of the power of expectations to bias judgment. This study definitely provided some such evidence, but: (1) a lot less than the study has often been cited as showing and (2) like Hastorf and Cantril (1954, see Chapter 2), there was considerably more evidence of social perceptual and judgmental reasonableness and accuracy than the study has ever been cited as showing.

First, let's briefly recap. Eight pseudopatients (who had no history, record, or diagnosis of mental illness) got themselves admitted to psychiatric hospitals complaining that they were hearing things (auditory hallucinations). Upon admission, they immediately ceased complaining of any symptoms of mental illness and acted as normally as possible under the (admittedly abnormal) conditions of the psychiatric hospital setting.

Next, let's evaluate whether the staff engaged in biased, error-prone, or inappropriate processing or judgments when they diagnosed these patients as schizophrenic. The second sentence in the prior paragraph should give you some reason for pause. *They were admitted complaining of auditory hallucinations.* Regularly hearing voices saying things such as "thud," "empty," and "hollow" is not remotely normal. If the pseudopatients had not been lying, such complaints would strongly suggest something seriously wrong somewhere.

As far as I can tell, therefore, an initial diagnosis of some form of schizophrenia does not seem to reflect any gross distortion on the part of the psychiatric staff. It was wrong, not because people with auditory hallucinations are no different from the rest of us, but because the pseudopatients were lying about their symptoms. It is pretty hard to hold the doctors culpable for not considering the possibility that the pseudopatients might be lying. Indeed, the doctors and staff probably arrived at as appropriate a diagnosis as possible.⁸

How rigidly resistant to change were the doctors' and staffs' expectations? Rosenhan's (1973) interpretation was that they were highly rigid. After all, none were diagnosed as sane—nearly all were released with a diagnosis of schizophrenia in remission. Rosenhan (1973, p. 252) seemed to think this was pretty telling: "... once labeled schizophrenic, the pseudopatient was stuck with that label. If the pseudopatient was to be discharged, he must naturally be 'in remission'; but he was not sane, nor, in the institution's view, had he ever been sane."

But let's focus on Rosenhan's actual results, rather than his interpretations. First, the average hospital stay was 19 days. In less than 3 weeks, on average, after admitting themselves with auditory hallucinations, the patients were released. How much shorter *should* the stay have been? If one of your relatives was having a psychopathic episode complete with hallucinations, would you want them released after a few days, or even a week or so, just because their symptoms had not (yet) reappeared? Of course, one patient was kept 52 days, which seems

awfully long, even to me (another example, in my opinion, of one of the main themes of this book—expectancy effects are usually small and occasionally large, and, as a general rule, people are rarely largely out of touch with social reality, although once in a while, we all can be pretty bizarre). On the other hand, this 52-day outlier also means that, on average, the other patients were only kept about 14 days. How much shorter *should* a stay be for someone who has been regularly hallucinating?

How about the diagnosis of “schizophrenia in remission”? Rosenhan’s take was that it showed that there was nothing these completely sane pseudopatients could do to convince the doctors that they were really sane. Doesn’t that show rigidity on the part of the doctors?

I am not so sure. Can a person ever have a psychotic episode and then be justifiably considered permanently sane? Regarding schizophrenia, the American Psychiatric Association’s Fourth Edition of the *Diagnostic and Statistical Manual of the Mental Disorders* (aka DSM-IV, 1994, p. 282) states: “Complete remission is probably not common in this disorder.” The DSM-IV also makes clear that the course of schizophrenia is usually very long term. Thus, in lay language, return to sanity is possible but unlikely, and, typically, schizophrenic symptoms recur over a long period of time. Given this, would it have been reasonable to pronounce a person sane after 2 weeks of acting reasonably following an episode of hallucinations? Probably not—such a conclusion would seem to be highly premature.

But there is even clearer evidence of high accuracy in Rosenhan’s (1973) data. Rosenhan performed a follow-up study at a hospital whose staff had heard about the pseudopatient study but who doubted that such things could occur at their own hospital. Rosenhan informed the doctors and staff at this hospital that some time during the upcoming 3 months, one or more pseudopatients would attempt to gain admission. Each staff member was asked to rate the likelihood that each new patient was actually a pseudopatient.

All staff with extended contact with or responsibility for patients, including attendants, nurses, psychiatrists, psychologists, and physicians, were asked to rate the new patients. Rosenhan (1973, p. 252) described his results this way: “Forty-one patients were alleged, with high confidence, to be pseudopatients by *at least one member* of the staff. Twenty-three were considered suspect by *at least one* psychiatrist” (emphasis mine).

The “at least one” phraseology is subtle, easily missed, and terribly important, at least with respect to figuring out how error-prone the staff were. It does not mean “all” or “most” staff identified 41 of 193 patients as sane. To be included in the 41, just *one* staff member had to suspect a particular person of being a pseudopatient. One out of how many? No one knows, because Rosenhan did not report how many staff members there were.

Despite this, we can make some plausible assumptions and approximate how accurate and inaccurate the staff and psychiatrists were. Rosenhan used the plural form when referring to each type of staff member. So let’s use some extremely conservative numbers. Let’s assume that there were only two nurses, two psychiatrists, two psychologists, and two physicians (larger numbers mean that the doctors and staff were even more accurate than indicated by these conservative estimates). That is a total of 10 judges. Each made 191 judgments of sanity (one for each of the 191 newly admitted patients). That is 1,910 judgments.

My guess is that Rosenhan used the “at least one” staff member phrasing because it was often the case that, for the 41 suspected pseudopatients, it was generally one and only one staff member so suspecting. Had it been *all* of the staff members, or even *most* of the staff

members, Rosenhan probably would have told us. If it is not “most” then the highest it could possibly be would be half the staff members. Again, this probably overestimates the number of staff members suspecting real patients of being pseudopatients, but let’s continue to sort this out in a manner that paints the doctors and staff Rosenhan studied in as inaccurate and biased manner as possible.⁹

In this analysis, I assume that all 10 staff members passed judgment on each incoming patient. Of course, it is possible that not all staff members passed judgment on each incoming patient. If that was the case, then, again, Rosenhan (1973) probably would have told us. This is because it would have strengthened his case. If pseudopatients were “identified” by a majority of those making these judgments, then the case would be very strong indeed for Rosenhan’s hypothesis that the sane are indistinguishable from the insane.

Given this, I assume that judgments of pseudopatients were not made by a majority of staff members. What would be the next best assumption for Rosenhan’s case? That half of the staff made these judgments of pseudopatients. This almost definitely overstates the total extent of error among the staff. Anything less than half, and the total amount of error on the part of staff will be even lower than my estimates. But let’s see where making this assumption leads us.

If, on average, half the staff (i.e., 5 of 10) identified each of the 41 people suspected of pseudopatients, that means there were 205 such judgments ($5 \text{ labelers} \times 41 \text{ patients} = 205$). By this analysis the hospital staff were correct nearly 90% of the time: 205 of 1,910 judgments were wrong, which means that 1,705 of 1,910 judgments were correct. In reality, accuracy was probably much higher, because there were probably more than 10 staff members (increasing the denominator, i.e., more than 1,910 judgments) and the proportion of suspects was probably less than half—in my example, less than 5 (decreasing the numerator, i.e., less than 205 pseudopatient judgments).

An approximate lower bound on bias, and an approximate upper bound on accuracy, can be estimated by assuming that *only* one staff member identified each of the 41 suspected pseudopatients. In that case, 41 of 1,910 judgments indicated suspected pseudopatients, indicating about 98% accuracy ($1910 - 41 = 1869$, which is the total of correct judgments, and $1869/1910 = 98\%$). Again, however, even this 98% figure could still be too low, if there were more than 10 staff members making judgments.

Given this analysis, let’s take the middle of my upper and lower bounds as a realistic guessimate of the accuracy of the staff at this hospital. That would place their accuracy at about 94%. Not perfect, but pretty good—especially 35 years ago, when the criteria for diagnosing mental illness were not as clearly spelled out as they are today.

Have Rosenhan’s (1973) results held up? Well, it depends on what the question means.

I consider two separable meanings: (1) Have Rosenhan’s conclusions regarding the powerful effects of psychiatric labels held up? and (2) Have Rosenhan’s results, which actually demonstrated some bias and a great deal of accuracy and reasonableness, held up?

Rosenhan (1973, p. 257) concluded that “We now know that we cannot distinguish insanity from sanity.” This conclusion has clearly not held up. Of course, I think even his own data failed to demonstrate this point. Although the doctors and staff may not have been perfectly accurate, for the most part, their diagnoses were reasonable and the average length of stay does not appear excessive. The results from the follow-up non-pseudopatient study indicated very high accuracy.

But there is another, even more important point. To suggest that a diagnostic label, such as schizophrenia, paranoid, or depressed, entirely or even mostly reflects the beliefs of the labeler is tantamount to claiming that differences between such people and others is entirely or mostly in the mind of the labeler, not in the behavior or experiences of the person so labeled. “Psychiatric diagnoses, in this view, are in the minds of observers and are not valid summaries of characteristics displayed by the observed” (Rosenhan, 1973, p. 251). This implies that there are little or no real differences between people identified as having such disorders and those not identified as having such disorders. This, in my opinion (intentionally or not), deeply denigrates and dismisses the depth of the psychological troubles experienced by people who typically come to be labeled, for example, schizophrenic, depressed, or obsessive-compulsive. After all, if there really is little difference between them and the rest of us, if the differences that we do perceive result entirely from the way diagnostic labels bias our judgments and perceptions, we do not need to provide them with any extra medical services, psychological services, care, or attention, do we?

If the problem is entirely with those doing the labeling, all we need to do is stop labeling them disturbed. That is, the way to treat depression, or bipolar disorder, is NOT to treat the patient; instead, we should treat those who have relationships with patients—their friends and family members. Does anyone really believe that we can cure schizophrenia or bipolar disorder by treating friends and family? Treating friends and family may actually have some benefits, but any perspective that implies that that will “cure” mental illness (in addition to being absurd on its face) seems to me to represent a recipe for increasing, not decreasing, social problems such as suicide, homelessness, murder, and spouse and child abuse.

Subsequent research. If the question “Have the results held up?” means, “Have Rosenhan’s results held up?” I think the answer is probably yes, keeping in mind that the actual results provided some evidence of biased perception and of considerable reasonableness and accuracy. The study has never been replicated, and probably could not be replicated, because admitting pseudopatients to hospitals would probably fail to meet modern ethical standards of the review boards that regulate research projects at most major research institutions.¹⁰

Nonetheless, considerable research on the extent to which psychological labels reflect real disorders versus perceptual distortions has been conducted. There is now abundant evidence that major psychiatric disorders are not mere diagnosis-confirming illusions in the minds of perceivers.

For example, there are real differences between children diagnosed with attention deficit/hyperactivity disorder and normal children—differences that are readily apparent even to other children (Harris, Milich, Corbitt, Hoover, & Brady, 1992). Furthermore, many of the symptoms of major mental illnesses, such as schizophrenia, depression, and bipolar disorder (manic depression in older parlance), can be reduced, sometimes dramatically, through medication (Bender, 1990; Ellison, 1989), thereby providing indirect evidence that there really was something physically/chemically/hormonally wrong somewhere.

In fact, the effectiveness of medication highlights the ultimate absurdity of a strong labeling perspective. To get concrete, this perspective would be compelled to predict that Prozac alleviates depression, not because it changes the emotions of those who take it, but because it changes the expectations of the perceivers who interact with those who take it. This analysis, which borders on silly, is my own extension of the labeling effects logic, so please do not hold Rosenhan (or anyone else) accountable for making such a claim. But it does not seem silly to

suggest that a labeling effects perspective assumes that psychotherapy drugs cannot possibly be effective by virtue of changing the emotions or mental state of the person taking them. Instead, the effectiveness of psychotherapeutic drugs, according to a labeling effects perspective, must derive entirely from their effects on *others' expectations* for the person taking the drugs.

Evidence regarding the accuracy versus biasing effects of diagnostic labels (learning disabled, emotionally disturbed, neurologically impaired, etc.) is more mixed. There is some evidence, for example, that physicians provide shoddier treatment to patients with psychological disorders (Graber et al., 2000), apparently because complaints from such patients are taken less seriously. This is a labeling effect.

The labeling issue comes up not only in psychiatric contexts but also in educational ones. Children struggling in school will often be sent for evaluation to a team of educators and/or psychologists and/or neurologists and will return with any of a host of diagnostic labels (e.g., emotionally disturbed, learning disabled, attention deficit disorder, neurologically impaired, etc.). Undoubtedly, many children are labeled appropriately, and this facilitates their receipt of appropriate attention and special programs. However, as of 30 years ago, as many as 40% of the children who received some label were misclassified (Ysseldyke, Algozzine, Shinn, & McGue, 1982). (I have not seen this research updated, so I do not know whether this figure would hold true today.) In addition, teachers who are more competent and more self-confident are *less* likely to refer children for the type of psychological evaluation that might lead to a label (Gersten, Walker, & Darch, 1988; Itskowitz, Abend, & Dimitrovsky, 1986; Meijer & Foster, 1988). Thus, although labels often aptly describe genuinely existing conditions, at least sometimes, the label reflects characteristics of the labelers as well as characteristics of the labelee.

Labels and labeling effects: Some bottom lines. The overall picture that emerges from this critical analysis of both Rosenhan's (1973) classic study and subsequent research is that labels can and do bias people's perceptions. However, such biasing effects generally tend to be small, although they can occasionally be quite large. Indeed, it is also clear that labels generally accurately reflect real characteristics of those labeled. That is why psychological labels are also called "diagnoses." People who are the targets of different psychological labels/diagnoses (schizophrenic, depressed, autistic, attention deficit/hyperactivity disorder) often are genuinely different than people without such labels/diagnoses. Furthermore, many of the people who interact with individuals labeled/diagnosed as psychologically disturbed (even when they are not aware of the label/diagnosis) successfully perceive differences between them and others without a label/diagnosis.

Whether there are any conditions under which psychiatric labels/diagnoses influence perception of targets more so than targets' actual behavior and characteristics influence perception remains unclear. It remains a challenge for those arguing for the existence of powerful labeling effects to empirically identify conditions under which such effects actually occur.

THE BIASING EFFECTS OF EXPECTATIONS ON JUDGMENTS AND EVALUATIONS: THE META-ANALYSES

My in-depth review and critical analysis of specific studies has been selective rather than comprehensive. This leaves open several possibilities that are damning to my claim here that

expectancy-confirming biases are generally small and fleeting, rather than powerful and pervasive. First, perhaps I have only selected studies that show small and fleeting effects. As I wrote in the introduction to this chapter, I have tried to do just the opposite—focus primarily on highly influential studies that have often been cited as testaments to the power and pervasiveness of expectancy effects. But perhaps I have, intentionally or unintentionally, biased the selection of studies to fit my conclusion.

Second, perhaps I have intentionally selected studies demonstrating bias that are particularly easy to shoot down because of obvious flaws. Metaphorically, perhaps I have set up a row of sitting ducks and then knocked them down. Third, perhaps I am simply unaware of numerous other studies, not included here, that have demonstrated more powerful and pervasive biases.

These three potentially damning criticisms of my review and conclusions themselves lead to empirically testable claims. First, if I have selectively included only studies demonstrating weak bias, then there should be many studies not included in my review that demonstrate strong bias. Second, if I have purposely selected sitting ducks, then there should be many other powerful, soaring birds out there that cannot be so easily shot down. And last, if I have missed lots of studies through ignorance, other reviewers, wiser and more knowledgeable than I, should have identified them.

These points (selective focus on weak bias, selective focus on easily criticized studies, lack of awareness of studies demonstrating more powerful effects) are all readily addressed through meta-analyses. Seven meta-analyses have examined the role of expectations in biasing judgments and evaluations. Together, they reviewed the results of 245 studies, including thousands and thousands of research participants. In one fell swoop, all of the potential damning problems associated with any “selectivity” in my review vanish because of the broad array of studies included in these meta-analyses.

So, what have the meta-analyses found? The second half of Table 6–1 has presented the results, but I review them here. As previously discussed, the Swim et al. (1989) meta-analysis of 119 studies of gender bias found an overall effect of $-.04$. In addition, Kunda and Thagard (1996) performed two meta-analyses: one of seven studies that examined the biasing effect of a stereotype in the absence of information about the personal characteristics of the target being judged and one of 40 studies that examined the biasing effect of a stereotype in the presence of information about the personal characteristics of the target being judged. These two effects were 0.25 and 0.19, respectively. Mazella and Feingold (1994) performed four meta-analyses of experimental studies examining the role of defendant social category in mock jurors’ verdicts. Effect sizes were 0.10 for attractiveness, 0.01 for race, 0.08 for social class, and 0.04 for sex, reflecting very slight tendencies to produce more innocent verdicts for attractive, White, middle class, and female defendants. The simple average of the effects obtained in these seven meta-analyses is 0.09. The weighted average (weighting the effect size by the number of studies included in each meta-analysis) is 0.07.

Translated into lay English, this means that expectations bias judgments, on average, about 5% to 10% of the time (see endnote 4). Or, put another way, on average, expectancies *fail to bias judgments* about 90% to 95% of the time. I conclude, therefore, that neither the handful of high-impact studies often cited as demonstrating powerful biases nor the broader, more general literature demonstrates that expectancies typically have very powerful effects on perception and judgment. Such biases are undoubtedly real—they occurred in many studies

reviewed here, and the meta-analyses consistently found evidence of bias. That being said, however, the only viable conclusion from this literature is that such biases are, in general, quite small.

EXPECTATIONS AND MEMORY

But what about memory? One source of the early enthusiasm for expectancy effects was that, not only did expectations seem to bias a wide range of perceptions, judgments, and evaluations, but also several dramatic studies suggested they seemed to exert a profound influence on memory. The next section, therefore, takes a closer and more critical look at some of the most dramatic and influential of those studies.

The saga of Cohen (1981). This was the study involving memory for the attributes and behaviors of the woman engaged in a dinner conversation with her husband, and in which half the time she was labeled as a librarian and half the time she was labeled as a waitress. Results showed that people consistently remembered 5% to 10% more stereotype-consistent than stereotype-inconsistent information about the woman.

I love (the relatively few) studies that provide clear, quantitative information bearing on the bias or self-fulfilling prophecy versus accuracy issue. Five percent to 10%—that should have sent off red flags (“Warning, warning: You are now entering a *weak* expectancy effect area”) when you first read about this study in Chapter 5. Would you describe that as a powerful expectancy effect? I wouldn’t—indeed, those numbers seem just about right in line with the typically small bias effects I have been claiming all along. Furthermore, Cohen (1981) also reported results regarding the accuracy of her perceivers’ memories. Across the two studies, accuracy levels were quite high—ranging from a low of 57% to a high of 88% and averaging about 75% in the first study and about 66% in the second study. High accuracy, small bias—this pattern should have a familiar ring to it by now.

One pattern of results from her second study was particularly relevant with respect to understanding the role of stereotypes in leading to bias versus accuracy. Cohen (1981) showed perceivers a portion of the same tape used in the first study. Half of the perceivers learned of the woman’s supposed occupation *before* viewing the tape; half learned of it *after* viewing the tape. In comparison to receiving the label after viewing the tape, when people received the label first, they more accurately remembered *both* stereotype-consistent *and* stereotype-inconsistent information. On average, they correctly remembered 70% of the target’s attributes (regardless of their degree of stereotype consistency) when they received the label first; they correctly remembered only about 63% of the target’s attributes when they received the label last. The upshot here, therefore, is that:

1. Although the label biased memory in such a manner as to favor stereotype-consistent information,
2. Having the label up-front also increased overall accuracy!

Why? Most likely, the label provides some sort of organizing scheme for perceivers, which facilitated their understanding and interpretation of both stereotype-consistent and stereotype-inconsistent attributes.

Thus, Cohen's (1981) research represents one of the very first demonstrations of a situation in which stereotypes led *both* to bias *and* to *increased* accuracy. This is extremely important, because, as shall be discussed in subsequent chapters, (1) studies that *only* test for bias (e.g., nearly all of the studies reviewed in Chapter 5) cannot possibly provide any information whatsoever about overall levels of bias relative to accuracy; (2) bias and accuracy are not necessarily mutually exclusive; (3) sometimes, bias may actually enhance accuracy; and (4) demonstrating that a stereotype *influences* social perception does not necessarily mean that it undermines the *accuracy* of social perception.

The saga of Snyder and Uranowitz (1978). This was the Betty K. study described in Chapter 5—the one in which Betty K. had a pretty mixed past, but which was misremembered in either a lesbian-consistent or heterosexual-consistent manner by perceivers (after “finding out” that she was either a lesbian or happily married). People not only more accurately remembered Betty's stereotype-consistent experiences but also reconstructed her past (misremembered her experiences) in a stereotype-consistent manner. Well done, dramatic study.

The problem is that each of two attempts to replicate the study failed (Bellezza & Bower, 1981; Clark & Woll, 1981). Each attempt actually included *two* experiments, so that in reality this constitutes *four* failed replications. Three of the four studies (both of Bellezza and Bower's and one of Clark and Woll's) used procedures highly similar to those used by Snyder and Uranowitz, and one (Clark and Woll's second study) used procedures *identical* to those used by Snyder and Uranowitz. None of these four experiments found any evidence of people reconstructing and misremembering Betty K.'s history in a stereotype-consistent manner (although Bellezza and Bower found that when people could not remember Betty K.'s history, they did rely on the sexuality label to help them make an educated guess about her past).

In discussing these failed replications, Snyder (1984) argued that their results were likely due to a fading of stereotypes about lesbians, and that such a fading was particularly likely in California (where, he claimed, they were conducted). This critique of the failed replications does not seem plausible for both theoretical and empirical reasons. There is almost no evidence of stereotypes dramatically changing in so rapid a time (3 years—1978 vs. 1981—the publication years of the original Betty K. study and the replications). The one exception, that is, the only time I am aware of such rapid stereotype change, was in time of unexpected war (e.g., American's stereotypes of the Japanese changed dramatically after December 7, 1941—see, e.g., Oakes, Haslam, & Turner, 1994, for a review of the conditions facilitating stereotype change). Furthermore, at least through the 1990s, abundant empirical evidence testifies to ongoing stereotyping of and prejudice toward gays and lesbians (e.g., Herek, 1993, 2000; Madon, 1997).

Nonetheless, Snyder's (1984) suggestion that “maybe people in California hold weaker stereotypes of lesbians” could have some validity. California has a reputation as a sort of a free-wheeling state, and the San Francisco area in particular is well-known for being a haven for gays and lesbians. So, perhaps, on average, Californians are more accepting of gays and lesbians than are people in other states. Even if true, however, this cannot account fully for these failures to replicate. Why? Because Bellezza and Bower's (1981) second failure to replicate was conducted in Ohio.

Expectations and memory: Some bottom lines. In addition to these specific failures to replicate, numerous studies showed *the exact opposite pattern*—that is, that stereotype- or

expectancy-*inconsistent* information was better remembered than was stereotype-consistent information (e.g., Bargh & Thein, 1985; Hastie & Kumar, 1979). This intellectual confusion, disarray, and disorder prompted the arrival of scholarly sheriffs, in the form of meta-analysts Stangor and McMillan (1992), to come in to try to clean up the town. Specifically, Stangor and McMillan performed a meta-analysis on 65 studies of the role of expectations in memory.

The main question was whether expectancy-consistent or expectancy-inconsistent information is best remembered. Overall—that is, averaging over all studies and all types of memory measures—the correlation between expectation and memory was 0.03.¹¹ Because of the large number of studies and subjects, this correlation was statistically significantly higher than zero. Whether such a low correlation has any practical significance is unclear. In practical terms, it means that, on average, expectations enhance memory for expectancy-consistent information about 1% to 2% of the time.

One of the main contributions, however, of Stangor and McMillan's (1992) meta-analysis was their examination of whether different results occurred for different types of memory measures. Specifically, they separated out results for free recall measures (what people spontaneously remember without prompting), recognition (when presented with some information about the target, identifying whether it was seen before), and response bias (the tendency to make expectancy-consistent guesses).¹²

Their results were quite interesting. On measures of free recall and recognition, there were small overall tendencies to better remember expectancy-*inconsistent* information ($r = -.08$ and $-.22$, respectively—the negative sign indicating that expectancy-inconsistent information was favored). On measures of response bias, there was a moderate tendency for people to make expectancy-consistent guesses ($r = .30$).

So, what does all this mean? The most likely explanation is that expectancy-inconsistent information is, by definition, surprising. Therefore, it really stands out and is more readily remembered. However, when people are not sure about whether they have received information about some target trait or behavior, they have to make an educated guess. And expectations guide those guesses, thereby favoring expectancy-consistent information.

This all makes sense to me. Some of my most memorable soccer coaching experiences occurred when the kids least expected to do so led our team to victory, or the kids most expected to lead us to victory made boneheaded mistakes that drove us down in defeat. But for those other games, the fuzzier ones that have mostly faded from my memory, if you ask me who did the scoring, I might, after scratching my head for a minute, say something like, "It was probably Rachel and Sarah" (the two high scorers on my team for years).

Returning to Stangor and McMillan's (1992) results, however, it is clear once again that expectancy effects were neither powerful nor pervasive. Even for response bias, the expectancy-based bias was only 0.3. This can be interpreted as a moderate effect of expectancies on educated guessing. It cannot, however, reasonably be characterized as demonstrating a powerful and pervasive effect of expectancies even on educated guessing (let alone memory in general). This 0.3 effect translates to expectancies changing educated guesses about 15% of the time (see endnote 4)—or, put another way, *not* changing educated guesses about 85% of the time. Furthermore, Stangor and McMillan's (1992) meta-analysis showed that *memory* is slightly better for expectancy-*inconsistent* information; when they cannot remember,

however, people do sometimes rely on expectancy-consistent educated guessing to help them out.

Conclusion: The Less Than Awesome Power of Expectations to Bias Perception, Judgment, and Memory

Do expectations lead to biases in judgment, perception, and memory? Yes, at least sometimes. But even the early research showed that such effects are, on average, neither inevitable nor anywhere near as powerful or pervasive as suggested in Chapter 5. As far as I can tell, the early research supported the following overall conclusions:

1. Expectations, including stereotypes and labels, sometimes bias perception, judgment, attribution, and memory.
2. Such effects are not only relatively small, on average, but also tend to be quite fragile, in the sense that seemingly small changes in experimental procedure, geography, type of dependent variable, or researcher often seem to lead such biases to mostly or completely evaporate and, sometimes, to completely reverse.
3. The studies from this early period that included an assessment of both expectancy effects and the accuracy or responsiveness to target individuating personal (rather than stereotypical) information (Baron et al., 1995; Cohen, 1981) showed that accuracy and individuating information effects are much larger than are the biases produced by expectations or stereotypes.

This brings us to accuracy. Just because bias tends to be small does not necessarily mean that accuracy tends to be high. Evaluating the accuracy question is simultaneously incredibly simple and dauntingly complex. Therefore, I discuss accuracy at length in the next three chapters.

Notes

1. Our decision process here may have been flawed—perhaps we relied *too much* on the practice results and too readily discarded our general knowledge/expectations about the differences between Amanda and Dori (e.g., Kahneman & Tversky, 1973; Locksley, Borgida, Brekke, & Hepburn, 1980). I say, so what? The issue here is not “look how great our decision processes are.” The issue here is a consideration of an everyday example in which, rather than rigidly clinging to expectations that powerfully biased our judgments, we readily discarded those expectations.

2. Meta-analysis is a highly sophisticated set of techniques used to summarize results from large numbers of studies. Conceptually, however, it can be reasonably simplified as referring to techniques for discovering the size of some effect averaging over many studies.

3. The authors here are listed alphabetically because order of authorship had not been decided, and never was decided because they ultimately did not publish the study.

4. Correlations have little intuitive meaning for most people but can usually be approximately translated into percentages, simply by dividing the correlation in half. So, a correlation (effect

size) of 0.20 means about 10% change, a correlation of 0.40 means about 20% change, and so on. The statistically inclined can find the mathematical formulas underlying this approximation in Rosenthal's (1984) discussion of the binomial effect size display.

5. Darley and Gross (1983) did not explicitly state the proportion of questions the girl answered correctly. Instead, they stated that she "... answered both easy and difficult questions correctly as well as incorrectly" (p. 23).

6. Whether or not this is a bias, however, depends on how closely people's social class-based expectations correspond to actual social class differences, an issue that is taken up *after* the next set of chapters on accuracy.

7. For the statistically inclined, random variation will produce a few extreme means purely by chance (remember that discussion of the sampling distribution of the mean in your first stats class?). Therefore, that one or two studies might produce an extreme mean (in either direction) is uninformative. In terms of drawing generalizations about effect sizes, therefore, I exclude the few analyses they conducted on fewer than five studies. Although they did report some results based on fewer than five studies, they, too, did not make very much of such small samples. Indeed, their overall conclusion was, "We found that the size of the difference in ratings between female and male target persons was extremely small . . ." (Swim et al., 1989, p. 419).

8. The point of this study was not to evaluate the validity of doctors' initial diagnoses. However, in the context of the overwhelming emphasis on error and bias in the social and cognitive literature (see Chapters 4 and 5, and especially the quotes toward the end of each chapter), and the extent to which this particular study has been cited as a testimony to the power of erroneous social beliefs, evaluating the validity of the doctors' and staff members' initial diagnoses itself is of some relevance and theoretical value. *My* point (whether or not it was Rosenhan's) is that their diagnoses seemed pretty darn accurate under the circumstances.

9. I am not purposely trying to make the doctors and staff that Rosenhan studied look bad. My point here, as will soon be shown, is merely that, even if we assume the worst, they did not look too bad!

10. This is not meant to cast aspersions on Rosenhan's ethics. Personally, I would not have moral qualms with such replication, but many university review boards, in my opinion, go too far in their attempts to ensure ethical research. But that is an issue beyond the scope of this book.

11. Stangor and McMillan (1992) actually reported memory effects in terms of d , which refers to the mean difference between groups divided by the standard deviation. ds , however, are readily translatable into correlations (see, e.g., Rosenthal, 1984, p. 25). However, when d is under 1, it approximately equals $2r$, where r is the correlation coefficient. Thus, a d of 0.06 corresponds to a correlation, r , of .03. Throughout this book, I almost always use rs or standardized regression coefficients, rather than other measures of effect sizes, primarily to use a single metric for evaluating effect sizes, and because it renders the experimental and meta-analytic research easily comparable to research reporting correlations or standardized regression coefficients.

12. Stangor & McMillan (1992) actually addressed many predictors of whether memory is biased in favor of expectancy-consistent or inconsistent information, and interested readers should refer to their paper for a more detailed and nuanced picture than I can possibly present here. Nonetheless, the broad patterns they found are derived from their Table 3, which is titled "Overall Effect Sizes and Homogeneity Tests."

4 Accuracy

CONTROVERSIES, CRITICISMS, CRITERIA,
COMPONENTS, AND COGNITIVE PROCESSES

This page intentionally left blank

10 Accuracy

HISTORICAL, POLITICAL, AND CONCEPTUAL OBJECTIONS

WHAT COULD BE a more basic or obvious purpose of social perception research than assessment of the accuracy of people's perceptions of one another? And what could be simpler? Although both questions are phrased rhetorically, in fact, accuracy was an all-but-dead topic within social psychology for roughly 30 years, from about 1955 to 1985. It turned out that the study of accuracy not only was less simple than it seemed but also was, in fact, a theoretical and political minefield. This chapter reviews, critically evaluates, and contests many of the reasons why social scientists have claimed that social perceptual accuracy is an unimportant, dangerous, or intractable topic.

This chapter is a polemic in the scholarly sense. Dictionary.com defines "polemic" as "a controversial argument, as one against some opinion, doctrine, etc." That is true of this chapter. I review many of the reasons why accuracy is a controversial topic in social psychology, I review the basis for considering some of those reasons to be politically motivated and to be at least a kindred of "doctrines," and I present some strong counterarguments to claims or suggestions that accuracy cannot or should not be studied.

A Lively Field Through the 1950s

Accuracy was a hot topic in the early years of social perception research. Researchers investigated a variety of accuracy-related questions, such as people's ability to predict their own and others' test performances, personality perception, and perceptions of agreement in attitudes and beliefs (Funder, 1987; Taft, 1955). At least two major theoretical perspectives were proposed at this time in which accuracy played an important role. Brunswik's (1952) lens model

was one of the first to explicitly consider accuracy a probabilistic concept (I am more accurate if I am right 85% of the time than if I am right 60% of the time) and to provide an analytic method for assessing perceptual accuracy. Brunswik's (1952) model suggested that accuracy and error in perception could be analyzed as a two-step process: (1) the probability that a person would use some stimulus cue and (2) the probability representing the cue's relationship to the attribute being judged. To present an oversimplified example, a social perceiver might judge targets' intelligence on the basis of their talking speed (faster equals smarter; thus, speed is the cue). Perceivers' judgments will be more accurate to the extent that talking faster actually reflects more intelligence and to the extent that they do not rely on stimulus cues unrelated to intelligence.

Kelly's (1955) theory of personal constructs incorporated accuracy in a very different manner. Kelly used the term "personal constructs" to refer to the idea that we all develop our own, somewhat shared, somewhat idiosyncratic systems of beliefs that we use for making sense of the world. We "construe" (interpret, make sense of, etc.) the world through the lens of our personal constructs. In this sense, Kelly's theory is compatible with Brunswik's. Brunswik's simply stated *that* perception could be considered a two-step process and provided some of the analytic means for assessing the process. Kelly's theory can be viewed as focusing more on how people decided which stimulus features of the environment were most important and how to interpret them.

Kelly's (1955) theory has often been viewed as emphasizing the subjective, phenomenological, "in the head" aspects of social perception and de-emphasizing the reality of things outside the perceivers' own constructs (e.g., Schneider, Hastorf, & Ellsworth, 1979; Wegner & Vallacher, 1977). Within social psychology, it has served as at least partial inspiration for research on implicit personality theory (people's beliefs about which personality traits go together with which other personality traits—e.g., Rosenberg, 1977) and self-schemas (people's beliefs about their own attributes—Markus, 1977). The emphasis in these lines of research has been entirely on the subjective—on people's *beliefs* about others or themselves; it has rarely addressed the relationship between subjective beliefs and actual reality. As a result, Kelly's (1955) theory developed a reputation as emphasizing the importance of the subjective construal part of social perception, rather than the link of perception to objective social reality.

Such a reputation is undeserved. Kelly's (1955) driving metaphor, which he stated explicitly, was "people as naive scientists." Why scientist? The following extended quote captures Kelly's (1955, p. 5) fundamental assumptions (and many of mine, too!) in a nutshell:

"... The scientist's ultimate aim is to predict and control. This is a summary statement that psychologists frequently like to quote in characterizing their own aspirations. Yet, curiously enough, psychologists rarely credit the human subjects in their experiments with having similar aspirations. It is as though the psychologist were saying to himself,

'I, being a psychologist, and therefore a scientist, am performing this experiment in order to improve the prediction and control of certain human phenomena; but my subject, being merely a human organism, is obviously propelled by inexorable drives welling up within him, or else he is in gluttonous pursuit of sustenance and shelter.'

Now what would happen if we were to reopen the question of human motivation and use our long-range view of man to infer just what it is that sets the course of his endeavor? Would we see his centuried progress in terms of appetites, tissue needs, or sex impulses? Or might he, in this perspective, show a massive drift of quite a different sort? Might not the individual man, each in his own personal way, assume more of the stature of a scientist, ever seeking to predict and control the course of events with which he is involved? Would he not have his theories, test his hypotheses, and weigh his experimental evidence? And, if so, might not the differences between the theoretical viewpoints correspond to the differences between the theoretical viewpoints of different scientists?"

Kelly (1955) explicitly assumed that the universe really exists and that the purpose of personal constructs was to help make sense of it. He acknowledged that sometimes constructs provided a poor fit to the data, but he pointed out that (1) a poor construct was usually better than *no* construct, which would leave the world as an incomprehensible mass of data input, and (2) people will typically revise their constructs (much like scientists may revise their theories) when those constructs do a poor job of allowing people to predict events in their daily lives or achieve their desired goals. Indeed, much of the goal of therapy, for Kelly (1955), involved getting people to (1) change their useless, outdated, or dysfunctional constructs to constructs that more closely reflected social reality and (2) try out many different constructs in order to better evaluate which ones work best.

Bottom line: Through the 1950s, accuracy was a thriving area of empirical research and a central component of at least two major theories. But all this interest in accuracy was about to come to a screeching halt.

Speculations on the Death of Accuracy Research

Understanding why accuracy research all but died out in the 1950s is not just an intellectual exercise in the history of social science (although it is, in my opinion, interesting as such an exercise). It is crucially important for understanding and evaluating the validity of perspectives emphasizing the power and pervasiveness of bias and self-fulfilling prophecy. If accuracy was mostly ignored for 30 years because it is of trivial relevance, then it might deserve to be ignored. Accuracy, in my opinion, is not and never was of trivial importance. But if accuracy is important, then why was it ignored?

THE FIRST CATALYST: CRONBACH'S REVIEWS

Two major reviews by Cronbach (1955; Gage & Cronbach, 1955) identified a host of apparent difficulties and complexities in assessing accuracy. These reviews rendered most prior research uninterpretable by virtue of not having addressed these difficulties (Cronbach's analysis will be discussed in detail in Chapter 12). Cronbach did offer solutions, but they were presented in a dense and complex mathematical style that daunted many researchers, especially in the precomputer era of data analysis (see, e.g., Funder, 1987; Kenny, 1994; or Chapter 12 of this book). Many researchers have concluded that the field died largely because accuracy research

became seen as hopelessly polluted by the artifacts, methodological nuisances, and components first identified in Cronbach's reviews (Cline, 1964; Cook, 1979; Funder, 1987; Jones, 1985; Kenny, 1994; Schneider et al., 1979).

This, however, was not the only reason accuracy research died out (Gilbert, 1998). Some of those other reasons are discussed next.

THE SECOND CATALYST: THE RISE AND INTELLECTUAL IMPERIALISM OF COGNITIVE PROCESS RESEARCH

I use the term "intellectual imperialism" to refer to the occasional tendency in intellectual/scholarly circles to attempt not only to promote one's favorite theory, perspective, or methodology, but also to denigrate, dismiss, and, in effect, quash alternative theories, perspectives, or methodologies. Within American psychology, for example, behaviorism from the 1920s through the 1960s is one of the best examples of intellectual imperialism. Behaviorists often characterized researchers taking other (nonbehaviorist) approaches to psychology as "non-scientific" (see, e.g. Skinner, 1990). And, although other forms of psychology did not die out, behaviorism dominated American psychology for four decades. Although behaviorism undoubtedly provided major contributions to psychology, to the extent that the scientific study of intrapsychic phenomena (attitudes, self, decisions, beliefs, emotions, etc.) was dismissed, ridiculed, or suppressed, behaviorism also *impeded* progress in psychology.

The rise of cognitive process research and E. E. Jones's dismissal of accuracy research. After Cronbach's reviews, an emphasis on cognitive processes dominated work on social perception and social cognition for the next four decades and, as far as I can tell, is alive and well today—and for good reason. Identifying the cognitive processes by which people arrive at judgments, perceptions, decisions, etc., is crucial with respect to understanding why people do what they do and think what they think.

But active interest is not the same as intellectual imperialism, which requires an active attempt to dismiss or denigrate other types of research. E. E. Jones was one of the central figures of American social psychology from the 1960s through the 1990s, and he was also at the forefront of those arguing that not only was research on process important but also that research on accuracy was an intellectual dead end (see, e.g., quotes in Chapters 4 and 5; Jones, 1985, 1986, 1990).

Furthermore, E. E. Jones (1985, p. 56) himself has pointed out that a few high-prestige individuals have an unusually large influence on the topics studied by other social psychologists:

More than most areas of psychology, social psychology is a personalized subdiscipline. People are often more concerned with "what Arbuthnot is up to" than with the state of knowledge on a particular topic. Prestigious researchers can be very influential in elevating the perceived importance of a research topic or claiming it for social psychology.

And E. E. Jones (1985, 1986, 1990) was at the forefront of those dismissing the value and viability of accuracy research. In 1985, he published a chapter in the *Handbook of Social Psychology*, titled "Major Developments in Social Psychology during the Past Five Decades." Such *Handbook* chapters are, arguably, the single most important and influential repository of the accumulated social psychology research, knowledge, and wisdom. Typically, only

the most prestigious, productive, and influential researchers are invited to provide such chapters.

And here are some selected quotes from the section on accuracy (p. 87):

Despite the obvious importance to social psychology of knowledge about person perception processes, the development of such knowledge was delayed by a preoccupation with the accuracy of judgments about personality.¹ . . . How can such factors (as halo effects) be checked or partialled out? What kinds of judges make accurate raters? What kinds of people and what kinds of traits are easy to rate accurately?

The naivete of this early assessment research was ultimately exposed by Cronbach's elegant critique in 1955. Cronbach showed that accuracy criteria are elusive² and that the determinants of rating responses are psychometrically complex. Prior to this pivotal analysis, however, Asch solved the accuracy problem by by-passing it.

Thus, I do not think it is too strong to suggest that E. E. Jones, himself, as one of the most eminent social psychologists of his era, was instrumental in the suppression of accuracy research and the rise and intellectual imperialism of cognitive process research.

Social psychology's dismissal of accuracy research. Regardless of Jones's role per se, however, it is clear that social psychologists have long downplayed the importance and value of research on accuracy. Indeed, although he undoubtedly influenced many social psychologists' view of the accuracy issue, he also spoke for and reflected the existing view held by most of the field. Years before Jones wrote his dismissive comments regarding accuracy, Schneider et al. (1979, p. 224), in what was the most influential text on social perception for over a decade, summarized the field's attitude toward accuracy as follows:

The accuracy issue has all but faded from view in recent years. . . . On the other hand, in recent years, there has been a renewed interest in how, why, and in what circumstances people are inaccurate.³

These are, I think, well-respected, common, and mainstream social psychological writings.

Consistently, about three-quarters of my students conclude that social psychology indicates that people are fundamentally irrational. Consider the following quotes from three student papers:

First student, introductory sentence: "Through taking this class, I have come to the conclusion that people are, and have always been, primarily irrational."

Second student, introductory sentence: "People are not rational beings; rather they are rationalizing beings."

Third student, concluding sentence: "I guess that we are probably irrational and spend our lives trying to convince ourselves that we are rational."

Furthermore, when I interview applicants to our graduate program in social psychology, I also ask them this question. Nearly all students—and especially the most well-trained ones—say "irrational and biased." This answer is wrong, but providing it increases my support for admitting them. Indeed, I *look* for that answer, because it tells me they have a good familiarity with

many of the major ideas in social psychology and are capable of drawing broad, big-picture inferences from that familiarity. It is not their fault that this answer is wrong. It is the fault of the scholarship in the field. Intentionally or not, scholarship in social and cognitive psychology, and the other social sciences, often creates the impression that people are fundamentally irrational.

Nonetheless, the view that people are fundamentally irrational and biased is not justified, not because it is completely wrong, but because it goes too far. The core purpose of error and bias research has been to reveal the *cognitive processes* underlying social perception (e.g., Funder, 1987; Jussim, 1991; Krueger & Funder, 2004). Such research is superb at doing so. For example, showing that people “see” three dimensions in a two-dimensional picture can be construed as an error or bias (there really are two, not three dimensions), but it also can provide information into the processes of visual perception (how people rely on lines, size, and angles to “see” depth). This does not necessarily mean, however, that visual perception suffers some sort of deep and fundamental flaw.

Similarly, showing that people sometimes rely on a stereotype when judging another person can be construed as an error or bias *if* the situation is constructed such that group membership is unrelated to the attribute being judged (which is the case in most social psychological studies of how stereotypes influence perceptions of another person). Such research can provide insights into the cognitive processes of person perception (how and when people use stereotypes, instead of or in addition to person information, to arrive at a judgment). This does not necessarily mean, however, that such judgments suffer some deep and fundamental flaw. If the belief about the group (stereotype) is well-grounded in reality, as some people’s stereotypes are (e.g., Lee, Jussim, & McCauley, 1995; and Chapters 16 and 17 of this book), and if we have little opportunity to obtain clear information about the person being judged, we will be more accurate more often if we use than if we ignore that stereotype (Brodtt & Ross, 1998; Jussim, 1991; Chapter 18).

WHY THIS EMPHASIS ON THE PROCESSES OF SOCIAL PERCEPTION AND DE-EMPHASIS ON THE CONTENTS OF SOCIAL PERCEPTION?

Accuracy is fundamentally about the contents of social perception. The most basic accuracy questions are: (1) What does someone believe? and (2) Does his or her belief correspond to reality? Both are questions about content, not process. The social cognition tradition, however, has long devalued or at least de-emphasized the worth of asking such questions in favor of questions about processes—cognitive processes, judgmental processes, information processing, and the like.

Why this four-decade-long emphasis on process? This is hard to say, but I have two broad sets of suspects. The first set of suspects involves reasons *to* study cognitive processes. Understanding cognitive processes is important and, to the extent that some researchers became more interested in process than accuracy, this was not a bad thing. The second set of suspects involves reasons presented *not* to study accuracy. These reasons were, in my opinion, bad ones and represent the intellectually imperialistic component of the rise of the emphasis on cognitive process.

First, let’s consider reasons to study cognitive processes. One was that content (of beliefs, perceptions, expectations) may come and go, but those emphasizing process often believed

they were getting at some fundamental and enduring aspects of human psychology. *What* people perceive or believe may be nearly infinitely variable, but understanding *how* people perceive their world, draw inferences about other people, or apply their beliefs to new information tended to be seen as holding out the promise of providing universal principles regarding human psychology (e.g., Fiske & Taylor, 1984, 1991; Kahneman, Slovic, & Tversky, 1982; Nisbett & Ross, 1980).

Most psychologists, including myself, once subscribed to this view. Nonetheless, empirical/scientific work on cross-cultural psychology, historical analyses of human psychology, and social constructivist/postmodernist work in anthropology and sociology have all cast serious doubts regarding the validity of this assumption. It is now clear that human thought processes are heavily bound, influenced, and altered by the prevailing conditions of one's culture, history, place in the social structure (power, status, wealth, etc.), and immediate or local social conditions (e.g., Ahearn, 2001; Baumeister, 1987; Berger & Luckman, 1966; Cerulo, 1997; Danziger, 1997; Holland, 1997; Nisbett, Peng, Choi, & Norenzayan, 2001).

Second, even if not universal, understanding social perceptual processes would seem to have greater generality than identifying the content of people's beliefs. For example, knowing that you think Clinton was a good president is all well and good but tells me nothing about what Ramona thinks. No generalizability. But if we learn that people's ideologies and attitudes color their interpretations of the actions and policy stances of political candidates, we may have learned something, not only about a liberal's attitudes toward Clinton, not only about a conservative's attitudes toward whoever is president, but also about many liberals' and conservatives' attitudes toward politicians' at almost any level of office.

In addition, in the 1960s, cognitive psychology began sweeping away the intellectual imperialism of the old behaviorist tradition (from which mentalistic concepts, such as expectations, attitudes, beliefs, etc., were all banished—see, e.g., Neisser, 1967). Starting with the schema concept and the metaphor of “mind as computer,” it provided a slew of new intellectual and methodological tools for examining cognitive processes in general and information processing in particular. Social psychologists were quick to adopt these tools to explore the information-processing underpinnings and effects of beliefs, attitudes, expectations, and stereotypes. Cognitive psychology opened a new research door and social psychologists eagerly and justifiably jumped right through it.

In addition, the focus on the *processes* of social perception and social judgment eventually led to a huge social psychological emphasis on errors and biases. Why? Again, it is hard to say. Part of the reason may have been that identifying consistent, systematic errors and biases provides insights into the processes of human thought (Funder, 1987). Part of the reason may have been that error and bias were unusual and attention grabbing—“man bites dog” is much more interesting, and much more likely to get publicity, than “man walks dog,” even though “man walks dog” is a far more common occurrence. But by relentlessly focusing on bias (because it was more interesting to most people, including researchers), it eventually appeared to be the norm rather than the exception.

Part of the reason may have been that experiments were designed in such a manner as to more readily produce evidence of bias than of accuracy or rationality (Krueger & Funder, 2004).⁴ Part of the reason may have been that identifying systematic errors challenged the implicit arrogance of everyday life, where most of us walk around thinking we are pretty good at figuring stuff out. Part of the reason may have been that social psychologists recognized

their own tendencies to commit these errors and, like other new converts, wanted to tell the world to help save people from themselves. Part of it may have been a genuine belief that many of the world's wrongs might be righted if people learned how to think more clearly (e.g., Gilovich, 1991; Hastie & Dawes, 2001; Heath et al., 1994).

Regardless, it is clear that flawed, biased, and inept cognitive processes typically caught social and cognitive psychologists' imaginations and opened up whole subfields of research on social cognition, information processing, attribution, and the psychology of prediction and decision making (e.g., Gilovich, 1991; Kahneman et al., 1982; Nisbett & Ross, 1980). Do I exaggerate the extent of the social psychological emphasis on error and bias? I do not think so. Consider the partial list of mainstream and classic social psychological phenomena presented in Chapter 1, in Table 1-1.

Understanding how, when, and why people go astray is an appropriate field of study for psychologists. In general, enthusiasm for new ways of doing research or understanding people shows a healthy and thriving psychological science. But, metaphorically, does man really bite dog more often than man walks dog (i.e., do error and bias dominate over accuracy)? Maybe so, but the only way we will ever find out is by conducting *both* error/bias research *and* accuracy research. This is, of course, not likely if accuracy research is dismissed and denigrated.

I do, therefore, have deep reservations about the attempts to quash, denigrate, or dismiss accuracy research that have appeared periodically for over 50 years. It is one thing to highlight complexities, difficulties, or limitations in assessing accuracy; it is quite another to suggest that such research is somehow more problematic and less viable than other types of research. Of course, perhaps accuracy research really is characterized by problems so deep that it cannot or should not be conducted. Next, therefore, I consider and critically evaluate some of the most common objections to accuracy research.

Objections to Accuracy Research: Point and Counterpoint

POLITICAL OBJECTIONS: "ACCURACY RESEARCH CAN BE USED TO JUSTIFY INEQUALITY"

Politics seems to lead to objections to accuracy research in two different ways. First, people's political stances may lead them to be more likely to raise scientific concerns about research that they perceive as opposing their political positions than about research that supports their political stances (e.g., Lord, Ross, & Lepper, 1979).⁵ This will almost never be stated quite so explicitly. No researchers will ever state "I am a liberal (or 'I am a conservative'); because I find this research politically offensive, I am going to work extra hard to come up with intellectual arguments against it." Explicitly stating this as one's position would seriously undermine one's credibility.

Furthermore, people may often not even be aware of how their politics influences their reactions to research. Thus, political issues often remain implicit undercurrents underlying objections to accuracy research rather than explicit statements. Occasionally, however, although no one has ever presented their own motivations as being starkly political, some researchers have presented starkly political objections or criticisms of accuracy research. Both types of political objections are discussed below.

WHY DOES ACCURACY AROUSE POLITICAL MOTIVATIONS?

Accuracy runs against the grain of many social scientists' concern for helping alleviate social inequalities and injustices. Accuracy cannot explain or alleviate social problems. Demonstrating that people's sex stereotypes are accurate (Swim, 1994) or that people's racial stereotypes are accurate (McCauley & Stitt, 1978) does nothing to alleviate or explain injustices associated with sexism or racism.

Worse yet, demonstrating social perceptual accuracy can be viewed as not merely documenting high acumen in perceiving individual and group differences, but also implicitly reifying and justifying those differences. To characterize a belief that some kid is not too bright, is a klutz on the basketball court, or is socially inept as "accurate" has a feel of "blaming the victim." Blaming the victim is a bad thing to do—it means we have callously joined the oppressors and perpetrators of injustice.

If the belief is "accurate," then we cannot point to perceivers' errors, biases, misconceptions, egocentrism, or ethnocentrism as explanations for target difficulties. The unintelligent, unathletic, or socially awkward target, in these cases, really is flawed in some way. This is especially true if the negative belief is applied to large demographic groups (i.e., stereotypes). Acknowledging this is difficult and distasteful (Tetlock, 2002). People who publicly declare that two groups differ in some societally valued attribute (intelligence, motivation, propensity for alcoholism or crime, morality, etc.) run the risk of being accused of being an "ist" (racist, sexist, classist, etc.) or, at minimum, of holding beliefs that do little more than justify existing status and hierarchy arrangements (e.g., Jost & Banaji, 1994; Sidanius & Pratto, 1999).

In contrast, an emphasis on expectancy effects or other errors and biases (including but not restricted to prejudice) implies a benevolent and egalitarian concern with injustice. Such an emphasis suggests that so-called "real" differences between groups do not result from any actual attributes of members of those groups (their cultures, their religions, their histories, their social conditions, their geography, their practices, their politics, their genetic predispositions)—they result solely or primarily from the oppressive effects of others' self-fulfilling stereotypes, prejudices, and expectations. Furthermore, this perspective suggests that many differences alleged to be real are not real at all—they simply reflect the ists' own expectancy-confirming biases.

In addition, an emphasis on expectancy effects provides a clear villain—the holder of the false expectation. It also points to a relatively straightforward way to ameliorate some social inequities—change expectations, stereotypes, etc. In contrast, not only does accuracy seemingly justify inequality ("they have lower status because they are less skilled, competent, intelligent" and so on), but also its relevance to solving social problems is not as readily apparent.

If my belief that you are incompetent is *inaccurate*, all that you need to do is change my belief to ameliorate the problem. But if my belief is *accurate*, then changing the situation requires much more work—to make us more equal, we have to upgrade your actual competence. And doing so may require years of education, training programs, mentoring, and the like. This is *much* more labor intensive and is not typically under the purview of most social psychological activities. All this may help explain the relatively greater appeal to many social psychologists and other social scientists of self-fulfilling prophecy and bias compared to accuracy.

EXPLICIT POLITICAL RATIONALES FOR QUASHING ACCURACY RESEARCH

The idea that accuracy research, at best, does not help anyone and, at worst, can be used to justify social inequalities seems to me to be a strong undercurrent underlying many of the theoretical objections discussed later in this chapter. Sometimes, however, a political rationale for quashing or dismissing accuracy research has been made explicit. Consider the following:

As scientists concerned with improving the social condition, we must be wary of arguments that can be used to justify the use of stereotypes . . . we cannot allow a bigot to continue to use his or her stereotypes, even if those beliefs seem to them to be accurate. (Stangor, 1995, p. 288)

“Improving the social condition” risks becoming a manifestly political agenda. Just what constitutes “improving the social condition” is likely to vary widely, depending on the political ideology of the person to whom you talk (consider such discussions with, e.g., Lenin, Hitler, Churchill, Roosevelt, Reagan, Yassir Arafat, Pol Pot, Malcolm X, Mahatma Gandhi, Martin Luther King, Rush Limbaugh, or Mother Theresa). Furthermore, this idea seems to come dangerously close to suggesting that our job is not to discover or report new information about social psychological phenomena or to fairly test hypotheses—it is, instead, to edit, censor, or skew our research such that they pass some sort of subjective and political “improving the social condition” litmus test. Who, I wonder, will administer this test?

I reject the premise. When I conduct research, my main goal is to find stuff out. If what I find out fails to fit someone else’s political agenda—tough darts. Of course, social scientists do have an ethical responsibility to make good-faith attempts to ward off *misuse* of their research findings. Fortunately, I have yet to come across a situation in which research on accuracy was misused to achieve some nasty purpose. More to the point, identifying misuse is a very different endeavor than is improving the social condition.

This is not to disparage research that seeks to improve the human condition. Advances in technology and research on how to elevate student academic achievement or reduce poverty are constructive fields of endeavor. These examples, however, are reasons *to* conduct research; I do not see how “improving the social condition” could ever constitute a scientific goal that could be achieved by *not* conducting research.

Stangor’s (1995) claim has one other thinly masked political assumption. It can be read as implying that a proper role for social scientists is to monitor and control how people use their beliefs (“We cannot allow . . .”). This statement may have been intended primarily to express support and concern for people from historically oppressed groups. Nonetheless, the idea that social scientists (or, indeed, anyone!) should “not allow” people to hold or express views with which they disagree (even if we label them with symbolically charged terms such as “bigot,” “racist,” or “sexist”) is chillingly reminiscent of Orwell’s *1984*.⁶

Consider also the following from Fiske (1998, p. 381) (commenting on the McCauley et al. [1995] chapter on stereotype accuracy):

Moreover, they differ from the present review in their conclusions, which do not follow from their premises⁷: If two resumes are otherwise equivalent, it is permissible to use

stereotypes associated with group membership as a factor in hiring choice, if group membership has previously predicted success on the job. (In this they evidently disagree with U.S. civil rights law).

“Permissible”!? We (I am one of the “et al.”) made claims about *accuracy*, not about “permissibility.” These are clear *political* rationales for quashing accuracy research. Fiske does not criticize accuracy research as failing to demonstrate that relying on stereotypes increases accuracy of predictions. Instead, she criticizes such research on grounds of the “permissibility” of relying on stereotypes. People in power make decisions about permissibility. I have never seen permissibility presented as a criterion for establishing the accuracy of judgments, and it is not included as one in this book.

Once the permissibility criterion is established, however, it has liberated Fiske (1998) to *completely ignore* our claim that relying on stereotypes sometimes increases the accuracy of judgments. Instead, she relies on two classic logical fallacies to make her point. First, she makes the “red herring” type of argument—being that she does not even attempt to refute the point that relying on stereotypes sometimes does increase accuracy, she changes the subject—to the politics of permissibility. She claims we disagree with civil rights law, *even though our paper never discussed civil rights law or any other law*. This also takes advantage of a second classic logical fallacy: the ad hominem attack. After all, who disagrees with civil rights laws except bigots? And we can’t believe anything a bigot says, can we? By implicitly insinuating political issues into the topic, Fiske’s quote is a masterful piece of misdirection, denigration, and dismissal that is likely to appear compellingly convincing to many of those sympathetic to her goals. But it is irrelevant to helping figure out whether and when relying on a stereotype increases or reduces accuracy.

Regardless of who agrees or disagrees with civil rights law, this is a fundamentally political, not scientific, rationale for quashing accuracy work. Fiske (1998) presents neither evidence nor argument that relying on group membership necessarily reduces the accuracy of perceivers’ judgments of individual targets. I submit that that is because she cannot do so, although to fully understand why, you will have to read Chapter 18 (if you are generally interested in the issue of accuracy and inaccuracy in stereotypes, you should read Chapters 15 through 19).

IS ACCURACY ILLEGAL?

While we are on the topic of politics, the law, and accuracy, perhaps it is sometimes *illegal* for people to arrive at the most accurate judgments possible. This is not as absurd as it might sound. Even though cognitive ability tests (IQ tests, SATs, GREs, etc.) are usually the single best predictor of performance on a great many tasks and in a great many occupations (Chapter 11), courts have sometimes issued rulings that have seemed to prohibit their use because they lead to underrepresentation of certain racial or ethnic groups (see, e.g., Gottfredson, 1994). In other words, courts have sometimes prohibited use of the most accurate predictors of future performance. Indeed, the courts have sometimes yielded down mutually exclusive principles *even within the same ruling* (see, e.g., Dawes, 1994; Gottfredson, 1994; Sackett & Wilk, 1994, for detailed discussions of this issue) regarding the use of nondiscriminatory criteria in personnel selection. That is, some rulings have (1) required personnel selection to use clearly

job-relevant criteria, (2) prohibited the use of criteria that lead to a disparate impact against a particular demographic group, *and* (3) prohibited selection on the basis of demographic group membership.

Such a ruling is incoherent because the three principles are mutually exclusive (Dawes, 1994; Gottfredson, 1994; Sackett & Wilk, 1994). Because of past or present discrimination, the demographic groups supposedly protected by such rulings often really differ from other groups on the most relevant criteria. Therefore, members of the protected groups will suffer “disparate impact” if those most relevant criteria are used in making personnel decisions. The only way to overcome such disparate impact would be to explicitly consider group membership in selecting personnel! This, however, has been explicitly *prohibited* in many legal rulings! Legal decisions, which are fundamentally political and some of which are logically incoherent, do not seem to provide a sound basis for establishing scientific criteria for accuracy.

THE PERVERSIVE INFLUENCE OF POLITICS IN THE SOCIAL SCIENCES

As far as I can tell, social and political judgments have the potential to influence all sorts of research, especially in the social sciences. For example, liberal commentators criticized Herrnstein and Murray’s (1994) research and conclusions regarding intelligence as motivated and biased by a right-wing political agenda (Jacoby & Glauber, 1995), and at least one conservative psychologist has criticized much of the research conducted by psychologists as being motivated and biased by a left-wing political agenda (Redding, 2001). Accuracy is, of course, no exception. But nor is it fundamentally more political than most other aspects of social science research. Indeed, politics, not theory, seems to be the primary basis for considering stereotype accuracy research to be a “problem” while at the same time not considering research on, for example, automatic stereotyping, in-group favoritism, and memory biases to be “problems” (these areas have their own substantial degree of theoretical and empirical controversy and complexity, so that it would be difficult to make the case that stereotype accuracy research is more problematic than other areas on purely scientific grounds). Regardless, as in most social science fields, strong steps can be taken to considerably reduce political bias from entering into consideration of issues involving accuracy (Chapters 11 and 12; Funder, 1987, 1995; Kenny, 1994).

ACCURACY AS A TOOL IN THE ALLEVIATION OF SOCIAL PROBLEMS

In addition, *even if our sole purpose in life was to alleviate social problems*, wouldn’t we want to know whether people’s beliefs about groups and their individual members (i.e., stereotypes) are accurate? It seems to me that we would, for several reasons. First, if we think we are curing a social disease by changing people’s inaccurate, biasing, or misbegotten beliefs about groups, our efforts will be sorely misplaced if their perceptions of groups and individuals are already accurate! Second, if *some* beliefs are widely inaccurate and some are reasonably accurate (as is likely the case), *only* research directly and empirically assessing the accuracy of stereotypes could possibly tell us *which* beliefs need to be changed through social interventions.

Furthermore, we need to be able to assess and understand accuracy in order to improve the quality of people’s judgments and evaluations. *Only* by developing techniques for assessing

the accuracy of people's beliefs can we possibly determine their *inaccuracy*. And only *after* determining that some people hold highly inaccurate beliefs would it be reasonable to begin work on changing those beliefs. Work on changing inaccurate beliefs itself would only be useful if it was conducted *after* we knew how to lead people to arrive at more accurate judgments. Of course, there will be no way to assess our success at leading people to adopt more accurate beliefs, unless we have techniques for assessing accuracy! By understanding what leads people astray and what leads them to accurate judgments, we will be much more capable of harnessing those factors that lead to accurate judgments and, therefore, reduce social problems resulting from inaccurate beliefs.

THEORETICAL OBJECTIONS

Not all objections to accuracy research are fundamentally political. Although politics may motivate people to develop more detailed and articulate arguments against research they oppose than against research they support (Lord et al., 1979), (1) once articulated, those objections stand or fall on their own merits, and (2) some objections to accuracy research may not have any political roots. Next, therefore, I consider some of the most common substantive and theoretical objections to accuracy research.

“COGNITIVE PROCESSES ARE IMPORTANT, ERROR AND BIAS ARE IMPORTANT, BUT ACCURACY IS NOT”

This strong argument has been explicitly articulated by various social psychologists (Jones, 1985, 1986, 1990; Schneider et al., 1979; Stangor, 1995). Furthermore, it is implicit in the topics studied by most social psychologists—with vastly more research on process, error, and bias than on accuracy. It is not merely that cognitive processes are important, with which I agree, but that accuracy is unimportant, with which I disagree.

My own view is that this objection falls apart from its own internal contradictions. It is like saying, “Let’s not look at baseball players’ batting averages, home runs, or RBIs, but let’s just analyze their swings to determine who is a good player” (the swing, of course, is a major part of the process by which they hit the ball). By this criterion, a hitter with a great-looking swing who gets a hit in 2 of every 10 at-bats and hits a homerun once every hundred at bats (i.e., one with a .200 average and about 5 or 6 homers a year) would be considered a better hitter than one with an awkward-looking swing who gets 3 hits every 10 at-bats and hits a homerun every 10 at bats (i.e., one with a .300 average and about 60 homers each year).

By the same token, psychological research articles are filled with excellent experimental studies of cognitive processes that researchers interpret as suggesting that bias, error, and self-fulfilling prophecy are likely to be common in daily life (e.g., Chen & Bargh, 1997; Fiske & Neuberg, 1990; Fiske & Taylor, 1991; Gilbert, 1995; Kahneman et al., 1982; Nisbett & Ross, 1980; Stangor & McMillan, 1992—see also Chapters 4 and 5). But such generalizations are only justifiable by research that examines the accuracy of people’s judgments in real-world contexts, not in artificial or even realistic laboratory contexts. No matter how much researchers *think* the processes discovered in the lab should lead to bias and error, the only way to find out for sure would be by assessing the accuracy of real social perceptions—just as the only way to discern which baseball player hit better would be by evaluating their success at hitting

(not by simply evaluating their swings). A social perceiver whose beliefs closely correspond to social reality is accurate, regardless of the processes by which that perceiver arrived at those beliefs.

Consider researchers studying basic social perception processes who wish to conclude that they have uncovered cognitive processes likely to lead to inaccurate or biased judgments in daily life (many researchers studying expectancies and stereotypes—see Chapters 4 and 5). This perspective leads inexorably to a simple hypothesis: The same types of beliefs and judgments are likely to be inaccurate and biased in daily life. The only way to test this hypothesis would be to assess the accuracy of beliefs in daily life. If those beliefs are found to be inaccurate and biased, this perspective would have garnered considerable support.

But if those beliefs are found to be mostly accurate and unbiased, then some aspect of these researchers' preferred hypothesis would be disconfirmed. Either the cognitive processes they identified are not routinely used by people or they are used but do not lead to biased and inaccurate judgments. Either way, it seems like something we should know. Either way, it would definitely tell us that the *implications* emphasizing error and bias that we have drawn from the experimental process research are not justified. And either way, it seems like it could enrich our understanding of cognitive processes by suggesting conditions under which a process does (e.g., the lab) and does not (e.g., daily life) occur, or by suggesting conditions under which a process that occurs in both the lab and real life sometimes leads to inaccuracy (lab conditions) and sometimes to accuracy (daily life conditions).

Indeed, one profound source of resistance to accuracy research may be precisely that it removes from researchers the ability to speculate on the power and pervasiveness in daily life of errors and biases found in the lab. Who wants to write a concluding section along the lines of: "We discovered this bias under highly artificial, ambiguous, or stimulus poor conditions, and even though we may have identified some basic cognitive process, there is good reason to believe these errors and biases rarely occur in daily life" (list of citations to research demonstrating accuracy)? Who wants to write a court brief in a discrimination case making the nuanced claims that "(1) lab research has uncovered numerous ways in which stereotypes may bias judgments; (2) nonetheless, it has also shown that people judge others far more on the basis of clear individuating information, such as very strong or weak job performance, than on the basis of stereotypes; and (3) research shows that stereotypes are accurate more than it shows they are inaccurate?" Such conclusions are too balanced and complex to provide the type of clear and convincing argument needed to help convince a judge or jury in an antidiscrimination court case.

So one "problem" with accuracy research may be that it removes from researchers the opportunity to speculate about the power and pervasiveness of the biases and errors they have identified. Because no matter how many flawed or imperfect cognitive processes may be identified, if people's perceptions, judgments, and expectations end up pretty accurate in real life, testaments to the power of those flawed or imperfect processes will not be justified.

This is obviously true in the case of stereotypes. In Chapters 4 and 5, and indeed, throughout the social sciences, one can easily find statements attesting to the pervasive inaccuracy of social stereotypes (see also reviews by Ashmore & Del Boca, 1981; Brigham, 1971; or almost any social science textbook in which stereotypes are discussed). But how can such statements be made, if accuracy research cannot or should not be conducted? Either accuracy research is not worthwhile, and all claims about both accuracy *and* inaccuracy would thereby be

forbidden from social science discourse, *or* accuracy research is essential to evaluate the validity of testaments to the inaccuracy of social perception in general and stereotypes in particular.

ACCURACY OF EXPLANATIONS: “JUST BECAUSE YOU SHOW THAT SOME BELIEF ABOUT SOME PERSON OR GROUP IS CORRECT DOES NOT TELL US WHY OR HOW THE PERSON OR GROUP GOT THAT WAY”

I have received this comment numerous times when giving research talks on issues of accuracy, bias, and self-fulfilling prophecy and when casually discussing these issues with colleagues. In addition, this criticism of accuracy research is implicit in perspectives arguing that accuracy cannot be studied or is meaningless because social processes and phenomena (e.g., discrimination, poverty) create the differences that are perceived (e.g., Claire & Fiske, 1998; Jost & Banaji, 1994). I am convinced, therefore, that it is a fairly common objection or, as shall be discussed, misunderstanding of accuracy.

I have two completely separate reactions to the claim that “demonstrating accuracy does not explain how or why those being accurately perceived got that way”: (1) This claim is absolutely correct and (2) it absolutely fails to threaten or undermine the viability, importance, or informativeness of accuracy research. I will illustrate both of these points with a hypothetical example.

Let’s say that Ben believes Joe is hostile. This “objection” focusing on the accuracy of explanations leads to at least *four* different questions: (1) Is Ben right? (2) What is Ben’s explanation for Joe’s hostility? (3) If Joe is hostile, how did he get that way? and (4) Why does Ben believe Joe is hostile?

Providing an answer to one question provides no information about the others. For example, establishing that Ben is correct (Joe really is hostile) tells us nothing about how Ben explains Joe’s hostility. Maybe Ben is a bigot who thinks that Joe’s ethnicity makes him prone to hostility. Maybe Ben thinks Joe was mistreated as a child. Maybe Ben thinks Joe hates his job. Maybe Ben thinks Joe watches too many old Clint Eastwood movies.

Similarly, establishing that Ben is correct tells us nothing about how Joe became hostile. Maybe there are genes for hostility and Joe has them. Maybe he was abused as a child. Maybe he has a miserable job and an unhappy family. Maybe Joe *has* been watching too many old Clint Eastwood movies. One could attempt to establish the validity of Ben’s *explanation* for Joe’s hostility by comparing it to the “true” reasons for Joe’s hostility, if they could be uncovered. Doing so would probably be a difficult task, but whole bodies of research have addressed sources of hostile and aggressive behavior (e.g., virtually every social psychological textbook has an entire chapter devoted to explaining aggression), so it would not be impossible. The important point is that assessing the validity of Ben’s belief *that* Joe is hostile is simply a different endeavor than is assessing the validity of Ben’s *explanation* for Joe’s hostility. The fact that a particular study only focuses on assessing one type of accuracy does not somehow fatally flaw such research—it only means that although considerable information may be provided regarding one type of accuracy (e.g., accuracy in perception of a trait), no information may be provided about another type of accuracy (e.g., accuracy in the explanation for that trait).

Furthermore, none of this necessarily explains how or why Ben came to believe that Joe is hostile. Maybe Ben regularly projects his own high level of hostility onto other people, so

that he is generally wrong but just happened to be right in the case of Joe. Maybe Ben heard that Joe is hostile from a mutual acquaintance and this expectation colored Ben's perceptions of Joe's behavior. Or maybe Joe is frequently insulting and sarcastic to Ben.

Understanding how Ben came to believe Joe is hostile is a very interesting and important question. It is a *social and cognitive process* question, and process is important. But it is an entirely different question than the others. Indeed, it is not an accuracy question at all. The accuracy issue evaporates here, because we are no longer evaluating the validity of Ben's perceptions, expectations, or beliefs. We are now trying to determine how Ben came to his beliefs.

This analysis is equally applicable to evaluating the accuracy of people's beliefs about groups (stereotypes). Determining whether Lois's belief that Asian Americans earn higher incomes than other Americans is accurate provides no insight into (1) Lois's explanation for the inequality, (2) reasons for the income relation between Asian Americans and other groups, or (3) how Lois came to this belief.

This should be obvious. Establishing *that* something is true is very different from establishing *why* it might be true. Establishing the accuracy of a person's explanation for why he or she holds a belief is a different endeavor than establishing the accuracy of the belief. Understanding *how* a person arrived at some belief is different than establishing the validity of the belief. Often, however, psychologists and other social scientists differ in their explanations for why people differ in almost everything (e.g., health, wealth, personality, intelligence, income, etc.). In the absence of a clearly well-established scientific explanation for many social phenomena, it may indeed often be impossible to evaluate the validity of laypeople's explanations for those phenomena. This does not mean that laypeople are wrong—only that the validity of their explanations cannot be determined. Although this may sometimes prevent assessment of the accuracy of people's explanations for individual and group differences, it does not detract whatsoever from our ability to assess the accuracy of people's perceptions of the characteristics and behaviors of individuals and groups.

ACCURACY VERSUS SELF-FULFILLING PROPHECY: "IT IS NOT MEANINGFUL TO DISCUSS 'ACCURACY' IF WHAT IS BEING 'ACCURATELY PERCEIVED' DOES LITTLE MORE THAN REFLECT SELF-FULFILLING PROPHECIES"

This is actually a variant of the prior objection ("demonstrating accuracy does not indicate how the person or group accurately perceived got that way"). This objection, however, specifies a very particular "way" that those being perceived accurately got that way—self-fulfilling prophecies. Although the entire analysis in the preceding section applies here as well, I give it separate consideration because (1) numerous researchers have specifically stated or can be read as implying that accuracy is meaningless because that which is accurately perceived could result from self-fulfilling prophecies (Claire & Fiske, 1998; Jones, 1986, 1990; Jost & Banaji, 1994; Snyder, 1984); (2) I have also heard this one numerous times when giving research talks or simply having informal discussions with colleagues on issues of accuracy, inaccuracy, bias, and self-fulfilling prophecy; and (3) both accuracy and self-fulfilling prophecy involve a belief or expectation corresponding well with targets' outcomes so that the potential confounding of the two is particularly salient or obvious.

The logic underlying this objection seems to be the following: (1) We know that self-fulfilling prophecies occur; (2) therefore, we also know that at least sometimes differences

between targets reflect effects of self-fulfilling prophecies; (3) if differences that are perceived reflect self-fulfilling prophecies to some unknown degree, attributing “accuracy” to those perceptions is, at best, meaningless and, at worst, reifies differences produced through social processes.

There is a kernel of truth to this argument. The first two premises are indeed true. Self-fulfilling prophecies do indeed occur sometimes, and, at any point in time, the differences between targets may indeed reflect self-fulfilling prophecies to some extent. Thus, differences that are accurately perceived at some point in time may reflect effects of prior self-fulfilling prophecies.

Furthermore, the confounding of self-fulfilling prophecy and accuracy clearly would be a problem in any situation (e.g., daily life, lab research) where it was not possible to distinguish these two very different reasons for why a perceiver’s expectations might be confirmed. Simply showing that a perceiver’s belief corresponds well with targets’ actual attributes or behaviors, by itself, cannot distinguish accuracy from self-fulfilling prophecy. Additional methodological procedures are required (beyond merely demonstrating correspondence between perceiver beliefs and targets’ attributes) to distinguish accuracy from self-fulfilling prophecy. Fortunately, such procedures are well-established, are well-known, and have been highly utilized.

A wide array of methodological and statistical techniques exist for distinguishing accuracy from self-fulfilling prophecy. Researchers have developed a wide repertoire of techniques for distinguishing accuracy from self-fulfilling prophecy. One is to have people judge targets with whom they do not interact (e.g., by judging them from resumes, college records, photographs, etc.). People cannot possibly create self-fulfilling prophecies among targets with whom they do not interact. Therefore, by ruling out self-fulfilling prophecy, such a design might allow for an assessment of certain types of accuracy.

Alternatively, some methods rule out the possibility of accuracy. For example, experimentally inducing false perceiver expectations allows for a clear assessment of self-fulfilling prophecy (most of the self-fulfilling prophecy studies reviewed in Chapters 4, 6, 7, and 8 used this methodology). At the end of the study, there are either differences between targets or not. Expectations are either self-fulfilling or they are not. A lack of differences between experimental conditions, however, which would mean there was no self-fulfilling prophecy, does not demonstrate accuracy. Indeed, although such designs do a good job of assessing self-fulfilling prophecy, they do not permit an assessment of accuracy.

Other methods allow for the simultaneous assessment of accuracy and self-fulfilling prophecy. Although a detailed discussion of these methods is beyond the scope of this chapter, the core idea is simple: if perceivers’ expectations are self-fulfilling, they should predict changes in targets’ behavior or accomplishments over time. Theoretical models relying on sophisticated statistical techniques have been developed for distinguishing self-fulfilling prophecy from accuracy under naturalistic conditions (Jussim, 1991; Jussim & Eccles, 1995; Trouilloud, Sarrazin, Martinek, & Guillet, 2002; West & Anderson, 1976; Williams, 1976). Many of these are discussed in detail in Chapter 13.

“Prior self-fulfilling prophecies may influence that which is ‘accurately’ perceived.” “Aha!” exclaim the accuracy naysayers. “That does not solve the problem.” “Why not?” I ask innocently. “Because even if *this particular* perceiver did not cause differences between targets, such differences may still have resulted from *prior* self-fulfilling prophecies that occurred in

interactions with *other* perceivers!" the naysayers declare with (premature) finality. I calmly sit back in my chair, put my feet up on my desk, and say, "Close, but no cigar." Here is why.

Perceivers' expectations cannot possibly cause differences among targets with whom they have not interacted. The first key idea is that if a perceiver cannot possibly have caused differences among targets, self-fulfilling effects of that perceiver's expectations cannot account for those differences. If the same perceiver successfully judges those differences, there will be a high correspondence between that perceiver's judgments and targets' attributes. When a perceiver's judgments closely correspond to targets' attributes, and when we know that that same perceiver's expectations cannot possibly have caused those attributes, by what term shall we refer to this correspondence? I think there is only one viable answer: accuracy.

Consider Coach Smith, the head coach of a high school girls' basketball team. Coach Smith observes two new girls trying out for the team, Donna and Mary. Donna finishes the 40-yard dash in 5 seconds, runs 3 miles in 20 minutes, hits 90% of her foul shots, hits 50% of her jump shots, and averages one rebound every 3 minutes. Mary, in contrast, runs the 40 in 8 seconds, cannot complete the 3-mile run because she becomes sick to her stomach, hits 40% of her foul shots, hits 20% of her jump shots, and does not pull down a single rebound in practice.

Coach Smith concludes that Donna is a better basketball player. Is there anything wrong, inappropriate, unjustified, or vacuous about Coach Smith's evaluations? Is there anything inappropriate, unjustified, or vacuous about considering Coach Smith's evaluation to be accurate? Is Coach Smith "blaming the victim"? Is she reifying differences between Donna's and Mary's ethnic groups? I do not think so. Coach Smith's judgment is accurate, even if the difference between Donna and Mary resulted, in part, from prior self-fulfilling prophecies. Perhaps Donna's parents strongly encouraged her participation in athletics, whereas Mary's parents did not. Perhaps Donna had a great coach in fourth grade who inspired in her a commitment to athletics, and perhaps Mary had a coach who was insulting and obnoxious, and who discouraged her. All these potential self-fulfilling prophecy explanations do not change the fact that, here and now, Donna is a much better player than Mary, and believing anything else would be wrong perhaps to the point of being silly.

Thus, the antiaccuracy argument is half right. Prior self-fulfilling prophecies occurring in other social interactions could indeed create real differences between targets that emerge in subsequent interactions. This does not, however, support or justify the conclusion that it is therefore meaningless or misleading to even try to assess accuracy. Real differences, once created, are . . . real! And one is more accurate if one recognizes them than if one denies them.

Alternative explanations: Hypothetical self-fulfilling prophecy explanations for target characteristics do not constitute evidence. But this antiaccuracy, pro-self-fulfilling prophecy argument itself has a major weakness. Just because someone can develop hypothetical self-fulfilling prophecy scenarios does not provide a shred of evidence that they are true. Just because some studies demonstrate that self-fulfilling prophecies do sometimes occur does not mean that they necessarily explain any particular differences between any particular people.

The self-fulfilling prophecy explanation for differences between Donna and Mary *might* be true, but there are tons of non-self-fulfilling prophecy explanations that also might be true. Perhaps Donna developed greater athletic prowess (speed, reflexes, etc.) as a result of constantly defending herself from her older brother. Perhaps they had similar parental or coaching experiences in the past, but Donna just liked basketball more and/or worked at it

more rigorously. Perhaps Donna's genes gave her greater height, stronger muscles, and less body fat than did Mary's genes. Perhaps Donna just happened to end up hanging out with a bunch of friends interested in sports, did what they did, and became good at basketball, whereas Mary's friends tended to prefer spending their time on phone conversations and shopping at the mall. Those possibilities, too, could explain why, here and now, there is a basketball skill difference between Donna and Mary. And, absent data, they are all as viable as hypotheticals as is the self-fulfilling prophecy explanation.

Is the self-fulfilling prophecy rejection of accuracy even scientific? This brief consideration of alternative explanations raises a bigger scientific and logical flaw in this objection to accuracy research. The "you cannot assess accuracy because of prior self-fulfilling prophecies" is also unjustified because it is founded on an untestable assumption. The argument implicitly assumes that it is possible to ascertain the ultimate or total extent to which people's characteristics result from self-fulfilling prophecies. Doing so would require obtaining empirical data on all of a person's social interactions throughout his or her entire life. Why? Because with only limited data, the accuracy naysayers could always claim that it was some *other* (i.e., not assessed in the study) self-fulfilling experience that created a particular target's characteristics.

I doubt that ascertaining the total extent to which anyone's characteristics result from self-fulfilling prophecies is possible. If not, then the claim that characteristics of a person resulted from unassessed (e.g., present before the study began) and hypothetical self-fulfilling prophecies is nonfalsifiable, and I subscribe to Popper's (1959/1968) view that falsifiability is one hallmark of a scientific theory or hypothesis. When scientific methods are developed for assessing the extent to which self-fulfilling prophecies over one's entire life contributed to one's attributes and skills, the self-fulfilling prophecy explanation for accurately perceived differences will become a scientific question. Until that time, however, although this explanation may have a certain intuitive plausibility, and may even be "true" in some extrascientific sense, it has no scientific standing.

The most that current social scientific research can accomplish, whether laboratory experiment, quasiexperiment, survey, ethnographic study, or observational study, is to provide information about relations between interpersonal expectations and social reality within some bounded context. The research context is typically bounded by interactants and time. Specifically, studies can and do address the relationship between expectations and social reality with respect to one particular pair or group of interactants (teachers and students, college roommates, coaches and athletes, parents and children, etc.). A study that focuses on lab interactions between strangers provides no information about how the parents of those strangers affected them. A study that focuses on managers and employees provides no information about teacher expectation effects.

Similarly, studies are bounded by time. Lab interactions between strangers typically take about an hour. Such studies, of course, provide no empirical information about the accuracy of social perception or the occurrence of self-fulfilling prophecies outside of that hour. Teacher-student studies typically take place over a school year. Such studies provide no direct empirical evidence regarding self-fulfilling prophecies or accuracy prior to that school year.

There is, however, one other way that the claim that self-fulfilling prophecies account for any particular differences between two individuals could be scientific (please bear with the following nonsocial example). We could not and do not need to watch a tree grow for

400 years to know that it is 400 years old. Why? Because we can simply count the rings. Trees grow one ring per year. No one has ever discovered a tree that grows one ring a day or one ring every 10 years. Thus, until the day comes when such trees are discovered, humans do not need to live 400 years to know that a tree is 400 years old—all they have to do is count rings.

The implicit logic underlying the antiaccuracy self-fulfilling prophecy analysis would need to be the same to achieve scientific credibility. If we actually had evidence of truly powerful and pervasive self-fulfilling prophecies; if study after study, with few or no exceptions, showed that interpersonal expectations created large and enduring differences between people; and if study after study similarly showed that *no other* biological or social phenomena ever created those differences, then, eventually, it would not be necessary to document that self-fulfilling prophecies explain any particular individual differences. It would simply be a reasonable assumption.

As Chapters 3 and 6 through 9 demonstrated, however, the accumulated evidence shows that expectancy effects are typically weak, fleeting, and fragile, rather than powerful and pervasive. Thus, the claim that accuracy is meaningless because self-fulfilling prophecies create the differences that are “accurately” perceived receives a veneer of scientific credibility only because it reflects the widespread assumption that self-fulfilling prophecies are powerful and pervasive. This claim makes little sense without this assumption.

Unfortunately, however, as shown in Chapters 3 and 6 through 9, the assumption is false (see also Brophy, 1983; Jussim, 1991; Jussim, Eccles, & Madon, 1996; Rosenthal & Rubin, 1978). Therefore, one cannot just assume that differences between any two people or groups result from self-fulfilling prophecies. Such a claim, therefore, requires specific empirical justification (i.e., empirical evidence that self-fulfilling prophecies caused the differences among the particular targets being studied). Citation of a handful of dramatic self-fulfilling prophecy studies (such as those discussed in Chapter 4) does not constitute adequate justification for a *new* claim that self-fulfilling prophecies caused differences among a new set of targets precisely because so much research shows that expectancy effects are far from inevitable.

The confounding of impressions and predictions I: The perceiver is accurate even if self-fulfilling prophecies resulting from other perceivers' expectations did create target differences. But there is another equally important flaw in this objection to accuracy research. It fails to account for time and, specifically, to distinguish between predictions and impressions. If target behavior, accomplishments, etc., predate perceiver beliefs about the target, causality can only flow in one direction: from target behavior to perceiver beliefs. Those perceiver beliefs may indeed become self-fulfilling—but only with respect to future target behaviors. I refer to perceiver beliefs regarding prior target behaviors, accomplishments, etc., as “impressions” and to perceiver beliefs that might predict future target behaviors as “predictions.” I illustrate the importance of this distinction by returning to our basketball coach and players.

Even if *all* of the self-fulfilling prophecy explanations for the difference between Donna and Mary were true, it would not detract at all from the appropriateness of characterizing Coach Smith's perceptions as accurate. Coach Smith just met the two girls this season. Smith, therefore, could not possibly have caused the difference she observed between the two girls. Events next Wednesday cannot possibly cause anything to happen today. If Smith just met the girls today, she could not possibly have caused them to develop different levels of athletic prowess yesterday or last year or 5 years ago. Thus, *Coach Smith's perceptions* of the two girls

cannot possibly have been self-fulfilling. If Donna really does play basketball much better than does Mary, and we know that Coach Smith did not cause that difference, would Coach Smith's perceptions be most accurate if she:

1. Perceived Mary to be a better player,
2. Perceived Mary and Donna to be equally good, or
3. Perceived Donna to be a better player?

The obviousness of the answer to this question makes salient why it is so important that the time and interactant boundaries of research be kept clear when considering the meaning of research on accuracy and self-fulfilling prophecy. Thus, it is definitely possible that differences between, for example, students as they first enter grade 6 reflect prior self-fulfilling prophecies. However, (1) the possibility that this is true does not make it true, and (2) even if it is true, it is utterly irrelevant with respect to assessing the accuracy versus self-fulfilling effects of the expectations held by the sixth grade teacher who had never met these students prior to sixth grade.

The confounding of impressions and predictions II: The perceiver's impressions can be accurate even if self-fulfilling prophecies resulting from the same perceiver's expectations did create target differences. Again, the key issue here is time. If my expectations trigger a social interaction sequence such that I cause you to become a very pleasant person, those expectations (which came prior to the interaction) are self-fulfilling. But, once our interaction is over, how should I perceive you? Would I be most accurate if I perceived you as nasty, as neither nasty nor pleasant, or as pleasant? Again, the answer is obvious. A "problem" arises only when we fail to distinguish between impressions and predictions (keeping in mind that today's impression can become tomorrow's prediction).

Thus, the claim that accuracy cannot be studied because prior self-fulfilling prophecies might influence that which is accurately perceived includes a core falsehood ("accuracy cannot be studied") enveloped in a good and valid point ("prior self-fulfilling prophecies might influence that which is accurately perceived"). Prior self-fulfilling prophecies *might* have influenced that which is perceived, but it does not mean accuracy cannot be studied.

THE CRITERION "PROBLEM"

The criterion "problem" has been one of the most common objections appearing in the literature criticizing accuracy research (e.g., Fiske, 1998; Fiske & Taylor, 1991; Jones, 1985, 1990; Schneider et al., 1979; Stangor, 1995). It is so common that it has been known to evoke paroxysms of sweat, angst, and even self-flagellation from people engaged in actual accuracy research. Aren't the criteria for evaluating the validity of social beliefs so vague and fuzzy as to render attempts to assess accuracy meaningless?⁸ Measuring extrovertedness, laziness, or intelligence is not like measuring heat or distance, is it?

Before directly addressing the criteria issue, however, I feel compelled to point out the ironic double standard inherent in this criticism. Heavy social psychological criticisms regarding the criteria used to establish accuracy exist side-by-side with the almost complete absence of such criticisms regarding the criteria for establishing self-fulfilling prophecies. Why is this so flagrantly hypocritical?

The double standard. I have never seen criticisms of the criteria used to establish self-fulfilling prophecies that remotely resemble those leveled at accuracy research. I find this peculiarly ironic because, although the processes by which a perceiver's beliefs become true are different, the *criteria* for establishing their trueness must be identical. Social psychology has a long history of exposing bias by exposing double standards. For example, if identical work is evaluated more positively when performed by John than by Jane, we have revealed gender bias (see Chapters 5, 9, and 18 for more on this). In a similar vein, therefore, if criteria are evaluated positively when used to study self-fulfilling prophecies but the same criteria are evaluated negatively when used to study accuracy, we have revealed a scientific bias. This state of affairs, therefore, constitutes another piece of evidence demonstrating social psychology's bias in favor of bias. Criteria used for assessing a self-fulfilling prophecy (which belongs in the broad family of social biases) are allegedly unproblematic; identical criteria for accuracy are allegedly so flawed as to render accuracy inordinately difficult to study.

When assessing both self-fulfilling prophecies and accuracy, the question is, "To what extent does the expectation correspond to the outcome?" There is a difference in the *process*—in *how* the correspondence comes about: (1) With self-fulfilling prophecies, perceivers' expectations predict targets' outcomes because they *cause* targets' outcomes; (2) with accuracy, perceivers' expectations *predict, but do not cause* targets' outcomes. Evaluating whether an expectation causes or predicts but does not cause some outcome is a process and research design issue—it is *not* a criterion issue. The criteria for establishing whether any particular belief is true must be just as good (or bad) as *the same criteria*, regardless of whether it is used in self-fulfilling prophecy research or in accuracy research.

If the use of some particular criterion is a "problem" in accuracy research, then, presumably, it should be just as much of a "problem" in self-fulfilling prophecy research. Indeed, all of the above measures—standardized tests, judges' ratings, self-perceptions—have been used in *both* self-fulfilling prophecy research and accuracy research (self-fulfilling prophecies: e.g., Rosenthal & Jacobson, 1968a; Snyder, Tanke, & Berscheid, 1977; Swann & Ely, 1984; accuracy: Goldman & Lewis, 1977; Hoge & Butcher, 1984; Ryan, 1995). Are standardized tests criticized for being inappropriate bases for judgments of intelligence (e.g., Gould, 1978; Jones, 1996)? If so, then using them would invalidate *both* accuracy and self-fulfilling prophecy research that uses them. Alternatively, if they are seen as an *appropriate* basis for establishing self-fulfilling prophecy (e.g., Rosenthal & Jacobson, 1968a,b), then they should be equally appropriate for establishing accuracy.

If comparison of perceivers' beliefs to judges' ratings does not necessarily indicate accuracy (it only indicates agreement, which may not be tantamount to accuracy—e.g., Kruglanski, 1989), then nor can it indicate self-fulfilling prophecy. Or, if comparison of perceivers' beliefs to judges' ratings is a good criterion for establishing self-fulfilling prophecy (e.g., Chen & Bargh, 1997; Snyder et al., 1977; Word, Zanna, & Cooper, 1974), then it must be equally appropriate for establishing accuracy.

I am not suggesting that any criteria are perfect (the criterion issue will be taken up in more detail in Chapter 11). All have limitations. My only point here is to highlight the extraordinary double standard that has historically developed within social and personality psychology regarding the criteria used to evaluate self-fulfilling prophecies and accuracy. The criterion issue is almost always raised in critical evaluations, discussions, and reviews of

accuracy research (e.g., Fiske, 1998; Fiske & Taylor, 1991; Funder, 1987, 1995; Jones, 1985, 1990; Kenny, 1994; Kruglanski, 1989; Schneider et al., 1979). It has almost never been raised in evaluations, discussions, and reviews of self-fulfilling prophecy research, some of which are by these very same scholars (Darley & Fazio, 1980; Fiske & Neuberg, 1990; Jones, 1986, 1990; Jost & Kruglanski, 2002; Miller & Turnbull, 1986; Olson, Roese, & Zanna, 1996; Snyder, 1984, 1992; see also the research cited and quoted in Chapter 4).

"Why not?" I can almost hear you asking. I am not sure, but I can offer a few speculations.

Why the double standard? Most scientific research traditions often, though not always, greet new ideas, approaches, and methodologies with considerable skepticism and criticism. Thus, the surprising thing is not that accuracy research has met some criticism. The surprising things are (1) that such criticisms, especially regarding the criterion issue, have led many researchers to conclude that the whole area is so befuddled with complexities, difficulties, and ambiguities as to not be viable, and (2) the lack of criticism regarding criteria for establishing self-fulfilling prophecies. Next, therefore, I offer some speculations regarding this odd state of affairs.

Metaphorically, just as a new romantic infatuation may blind those involved to their partner's limitations, social psychology's early infatuation with self-fulfilling prophecies may have blinded researchers to some obvious limitations. That infatuation itself probably stemmed from at least two sources. The first was the theoretical zeitgeist of the 1970s and 1980s, which emphasized psychological processes, errors, and biases. Social psychologists' enthusiasm for bias and error may have led them to be less critical of self-fulfilling prophecy research than of accuracy research.

This double standard may also reflect an implicit effect of the political issues discussed earlier in this chapter. Social psychologists' social activism/social problems orientation may have (even if unintentionally) helped shield self-fulfilling prophecy research from the type of scrutiny commonly applied to accuracy research. People often more rigorously scrutinize research opposing their political views than they scrutinize research supporting their political views (Lord et al., 1979). Perhaps a similar phenomenon may help explain the greater critical scrutiny of the criteria for establishing accuracy than of the (identical) criteria for establishing self-fulfilling prophecy.

Regardless, the social sciences cannot have it both ways. Despite differences in the processes by which accuracy and self-fulfilling prophecy occur, establishing both accuracy and self-fulfilling prophecy requires establishing correspondence between a social belief and criteria. High correspondence means *either* accuracy or self-fulfilling prophecy (or some combination of both) have occurred; low correspondence means neither has occurred (at least not very much). Once high correspondence is established, additional methodological procedures are required to distinguish accuracy from self-fulfilling prophecy—but both require high correspondence between belief and criteria. It cannot be tortuously difficult or impossible to identify criteria for establishing accuracy unless it is equally tortuously difficult or impossible to identify criteria for establishing self-fulfilling prophecy. Conversely, it cannot possibly be unproblematic to identify criteria for establishing self-fulfilling prophecy unless it is equally unproblematic to identify criteria for establishing accuracy.

In short, any social psychological perspective that lambasts accuracy research for lacking criteria but extols the value and importance of self-fulfilling prophecy research (Claire & Fiske, 1998; Fiske, 1998; Jones, 1986, 1990) is logically incoherent.

Conclusion

Within social psychology, accuracy research has had a turbulent and controversial history, which likely explains why such little research on accuracy was performed from about 1955 to 1985. This, in turn, may help explain the absence of accuracy from most major reviews of interpersonal expectancies (e.g., Claire & Fiske, 1998; Darley & Fazio, 1980; Jones, 1986, 1990; Miller & Turnbull, 1986; Synder, 1984, 1992), although some reviews starting in the late 1990s began to seriously grapple with accuracy issues (Olson et al., 1996; Snyder & Stukas, 1998). The combination of the banishment of accuracy with the infatuation with bias helps explain the current extraordinary state of affairs: Even well-intentioned, balanced, even-handed scholars often find themselves compelled to conclude that the last four decades of research in social psychology overwhelmingly demonstrates the flaws, shortcomings, irrationalities, and biases of human social judgment and social perception. The journals are filled with studies extolling bias (even studies that often provided more evidence of accuracy—see Chapters 2 and 6 through 9), not because bias dominates over accuracy, but because the *study* of bias dominated over the *study* of accuracy.

In this chapter, I have suggested that many of the claims raised in the context of objections to accuracy research have considerable validity, in the sense that much of the content of the criticisms may be true and they often raise interesting and important questions. Accuracy in perceptions of an attribute really is different than accuracy in explanations for that attribute. Prior self-fulfilling prophecies might explain some differences between targets that are accurately perceived. And the criterion issue is an important one.

I have also argued, however, that despite whatever validity they might have, none of the criticisms warranted abandoning accuracy research completely or the conclusion that accuracy is so hopelessly confounded with social processes and self-fulfilling prophecies as to render the construct meaningless. Nonetheless, claiming that the arguments against accuracy research are themselves flawed does not indicate how accuracy research can be conducted. That issue is taken up in the next two chapters.

Notes

1. This seems to assume that research on accuracy does not provide information on process. Such an assumption is unwarranted (see Chapters 12 through 14 and 18).
2. Although Jones was a brilliant and careful scholar, I find this claim hard to understand because Cronbach's 1955 critique did not address the criteria issue at all. I suspect this represented Jones's, not Cronbach's, view—but see Chapter 12 for a detailed discussion of Cronbach's view.
3. Such a claim seems, on its face, incoherent. How can one study inaccuracy without studying accuracy? Schneider et al. (1979), it should be noted, were summarizing widely held views within social psychology, so I am *not* suggesting that Schneider et al. necessarily ascribed to this view.
4. Bias was typically assessed by examining whether people in one group (Group A) viewed (interpreted, remembered, evaluated) some stimulus *differently* than did people in another group (Group B). Groups A and B might be given different stereotypes, different expectations, different personality information, etc., about some stimulus (event, person, group, etc.). Nearly all of the research discussed in Chapters 2 through 9 fits here. All such research is skewed toward finding bias.

This is because statistically significant differences between groups can *only* occur if people are biased. Lack of bias would only produce nonsignificant differences between groups. Nonsignificant differences (1) are unlikely to be published in scholarly journals, thereby leading to a publication bias against studies finding no bias, and (2) are theoretically uninterpretable (as everyone is taught in one's first statistics class, it is only possible to reject the null hypothesis; it is never possible to accept the null hypothesis). Thus, because rationality and accuracy could only be reflected in nonsignificant differences in such designs, and because nonsignificant differences are theoretically uninterpretable, such designs rendered it impossible to find evidence of rationality and accuracy (see Krueger & Funder, 2004, for a more detailed exposition of this issue).

5. How can I cite a "bias" study in a chapter arguing that social psychologists overstate error and bias and have inappropriately dismissed accuracy? Lots of reasons: (1) I never denied that biases occur; (2) this book is about expectancies, not politics; (3) people's political positions influencing their evaluations of scientific research is a different phenomenon than people's expectations influencing their behavior toward or judgments of other individuals; so that (4) social psychologists' politics may well influence their evaluations of accuracy research more so than laypeople's expectations influence their evaluations of other people. One reason this may seem odd is that I appear to be arguing that social psychologists, who are highly trained experts, are more biased than are untrained laypeople. I am making no such argument. I am not suggesting, for example, that social psychologists' politics influence their evaluations of research more so than laypeople's politics influence their evaluations. I am suggesting, however, that social psychologists' politics may have unduly influenced their interpretations of research on error, bias, and accuracy.

6. I doubt that Stangor's views reflect quite as much authoritarianism as this quote seems to express. Although the quote is both accurate and in context, perhaps Stangor meant something more benevolent, such as "we cannot allow a bigot's views to go unchallenged." How do we challenge a bigot's views? One powerful tool is by showing that they are inaccurate! But to do that implies that it is possible to hold an accurate belief. And that can only be accomplished if we can and have obtained scientific evidence on accuracy!

7. I respectfully disagree with Fiske's claim here that the conclusions of McCauley et al. (1995) do not follow from their premises. Indeed, Fiske never specifically stated which particular conclusion failed to follow from which particular premise. I suspect this is because there was no logical error to be identified, but to decide for yourself, you should read that chapter!

8. Establishing the accuracy of some social belief requires some sort of standard of comparison. "Criteria" here and in the next chapter *does not* refer to the overall scientific and methodological procedures and processes used for assessing accuracy—it refers only to the question: "What outcome or measure shall we use as a standard against which to evaluate the degree of accuracy and inaccuracy in some individual's or group's belief?" To get concrete, if Fred believes that Joe is rich and hostile, how will we measure Joe's wealth and hostility? Income tax returns? Value of Joe's home? Joe's scores on a hostility scale? Co-worker assessments of Joe? The criteria issue refers to our choice of standard against which to assess the validity of perceivers' beliefs.

11 Accuracy CRITERIA

HOW DO WE know what we know? This chapter, and indeed this book, will not address this question, at least not directly. It is too broadly philosophical, having roots at least as deep as Descartes' "I think therefore I am." But it has two aspects that will be addressed in this chapter, one indirectly, the other directly.

One aspect can be phrased as "How do we come to know what we know?" This question is primarily a process question, not an accuracy question, and evaluating the accuracy of some of the processes involved in social perception will be addressed in Chapter 12. One such process, however, is partially addressed here. Social reality causing social beliefs can represent one type of accuracy. Showing that some particular social belief was soundly based on some aspect of social reality requires, in part, establishing correspondence between belief and reality. Precisely what reality? And how shall we assess it? Establishing correspondence between belief and reality returns us to the criterion issue. Thus, the criterion issue is intimately interwoven with some aspects of the "How do we come to know what we know?" question.

The second aspect, which will be addressed directly, can be phrased as "How can we evaluate the validity of what we think we know?" This *is* the criterion issue.

If The Weather Channel predicts 90% chance of thunderstorms tomorrow, how do we know if they are right? This is easy: If thunderstorms occur 90% of the time they claim a 90% chance of thunderstorms, they are right.

Establishing criteria for evaluating the accuracy of beliefs regarding physical events, such as rain, speed, size, etc., is easy. But the criteria for evaluating beliefs about other people often are not as clear and objective. How can we evaluate laziness, intelligence, courage, or friendliness? One can look outside one's window and see if it is raining; one cannot literally look at an individual to see if it is raining on their soul (e.g., that they are seriously depressed).

So, then, shouldn't we just throw in the towel? Aren't social attributes so ambiguous that it is either impossible to assess accuracy in perceiving them or at least so difficult that the effort is not worthwhile? Isn't finding criteria against which one can establish the accuracy of most social beliefs a Quixotic quest for a nonexistent Holy Grail? Some of the most famous and influential social psychologists of all time have argued that the answer is, essentially, a resounding "yes" (Fiske, 1998; Jones, 1985, 1986, 1990; Kruglanski, 1989; Schneider, Hastorf, & Ellsworth, 1979).

I respectfully disagree. Here is why.

Theoretical Perspectives on Criteria

The appropriate criteria for establishing accuracy will, of course, depend on what one thinks accuracy is. Next, therefore, I present an overview of three very different general perspectives on accuracy. I also explain why I only see one of them as providing a scientifically tenable approach to accuracy.

PROBABILISTIC REALISM

The main ideas of probabilistic realism, which I adopt throughout this entire book, are that there is an objective reality out there that, flawed and imperfect though we may be, we can eventually come to know or understand, at least much of the time. I use the term "probabilistic realism" as a somewhat simpler name for what has been called "critical realism" and "pan-critical rationalism"—this is essentially the same approach described by Funder (1995), which itself was heavily influenced by mainstream psychological approaches to construct validity (e.g., Cook & Campbell, 1979; Cronbach & Meehl, 1955). Most scientifically-oriented researchers implicitly adopt this perspective nearly all of the time (at least in their research).

Nonetheless, precisely because this perspective often remains *implicit*, many researchers, especially those who dismiss or denigrate accuracy research (see Chapter 10), may not even be aware that they adopt this perspective. It is important, therefore, to make its main ideas explicit. Next, therefore, I describe both the "realism" and "probabilistic" aspects of probabilistic realism, and then briefly describe what accuracy means in this context.

Realism

"Realism" refers to the idea that there is an objective reality out there that is independent of social perception. Indeed, few social scientists of any stripe, except the most radical of social constructivists, deny the existence of such a reality (constructivists will be discussed later in this chapter). Even the three-decade exile (roughly 1955–1985) of accuracy research within social psychology occurred because of heightened interest in bias and recognition of genuine complexities in studying accuracy—not because most researchers explicitly argued there was no reality out there (see Chapter 10). Indeed, the idea that there is an objective social reality out there that influences social perception and constrains the potential for bias is quite explicit in many theoretical perspectives within psychology (e.g., Allport, 1955; Brunswik, 1952; Festinger, 1957; Gibson, 1979; Jussim, 1991; Kelly, 1955; Kunda, 1990; McArthur & Baron, 1983). The "realism" part of probabilistic realism reflects this assumption.

Social Reality and Social Beliefs Are Often Inherently Probabilistic

I use the term “probabilistic” to capture three different aspects of accuracy. First, it means that most criteria are probabilistic, not definitive. A student with a high IQ is *likely* to do well in school, but there is no guarantee. The winners of last year’s U.S. Open tennis tournaments will probably do well again this year, but, not only is there no guarantee that they will win again, but also they could get knocked out in the first round (although they are less likely to do so than most other players). If Nepalese are more courageous than other people, this does not mean that there are no cowardly people from Nepal.

Second, “probabilistic” captures the idea that many social beliefs themselves are inherently probabilistic. The belief that Michael Jordan was the best basketball player of the 1990s *does not* require believing that his teams would win every game they played. Although people rarely phrase their beliefs in explicitly probabilistic terms, this belief can be interpreted as meaning something like “all other things being equal, having Jordan on your team will enhance your chance of winning more than having any other player on your team.”

Stereotypic beliefs, too, are usually inherently probabilistic (see also Krueger, 1996; McCauley, Stitt, & Segal, 1980). A belief that Asians are wealthier than other people does not necessarily mean that the belief holder expects all Asians to be fabulously rich or denies the existence of a single impoverished Asian—only that, on average, they are richer than other people. Of course, even absolutist beliefs can be viewed as probabilities. The belief that all Englishmen are dignified can be translated into the belief that 100% of Englishmen are dignified.

Third, I use the term “probabilistic” loosely to reflect the inherently quantitative, rather than absolutist, nature of accuracy. That is, if Jorge expects John to be late for all of their meetings, and John is late only 95% of the time, Jorge is still pretty darn accurate—certainly far more accurate than had he expected John to generally be on time. Accuracy is rarely all or none; it is usually a matter of degree.

Accuracy

In this context, social perceptual accuracy is correspondence between perceivers’ beliefs (expectations, perceptions, judgments, etc.) about one or more target people and what those target people are actually like, independent of perceivers’ influence on them. More correspondence without influence, more accuracy.

Let’s unpack this definition. It has three main ideas: correspondence, what people are actually like, and independent of influence. “Correspondence” is easy. If I predict Bella will receive an A on her next math test and she does, my belief corresponds well with the outcome. Similarly, near misses involve closer correspondence than wildly inaccurate misses. If I predicted an A for Bella and she receives a B+, I am still more accurate than you if you predicted that she would receive a C.

“Independent of influence” is conceptually fairly easy, too. It means that I cannot have caused your outcome. If I predict that you will become the team’s best player and I cause you to become the best player (e.g., by giving you extra time and attention, being extra supportive and encouraging to you but not to other players, etc.), this is a self-fulfilling prophecy, not accuracy. Although one could interpret a self-fulfilling prophecy as a type of accuracy (see Swann, 1984), I draw a hard conceptual distinction between the two. For my belief to be

accurate, it must correspond to your behavior or attributes without having caused them (although many social scientists have implied that separating out accuracy from self-fulfilling prophecy is unimaginably difficult, they are wrong; unfortunately, you have to have read Chapter 10 to understand why they are wrong).

The phrase “actually like” is deceptively simple. Because this phrase implies a host of assumptions that other people may or may not share, I make them explicit here. First, I am now drawing on my realism assumption. That is, I assume that people could have some characteristics that are independent of my judgments of those characteristics. Second, how to identify what those characteristics are, independent of subjective interpersonal judgment, is, in essence, the criterion issue. This issue will be addressed immediately after briefly discussing the other two main theoretical approaches that might be seen by some as relevant to the accuracy issue.

FUNCTIONAL PERSPECTIVES

Functional perspectives emphasize the psychological phenomena and processes that help people function well. The central issue for functional perspectives involves determining how well some psychological phenomenon helps people get through the day, be happy, or accomplish their goals (e.g., Snyder, 1992; Swann, 1984). For functional perspectives, therefore, evaluating accuracy means determining how well a belief, expectation, stereotype, schema, etc., helps perceivers get what they want. Social beliefs that help people accomplish their goals are more accurate than beliefs that do not help people accomplish their goals (McArthur & Baron, 1983; Swann, 1984). This perspective requires (1) identifying what beliefs people hold and (2) determining whether those beliefs help or hinder them in obtaining desired outcomes.

I do not adopt a functional perspective. In my opinion, by focusing on perceivers’ goals, happiness, etc., such perspectives can avoid the issue of whether perceivers’ beliefs correspond to anything remotely resembling an independent and objective social reality. Consider one group of people who would like to exterminate another group of people. If believing that the target group is immoral, malicious, and subhuman helps the first group engage in genocide, it seems that those beliefs would have to be considered “accurate” within a functional perspective.

This may be internally consistent with the logic of a functional perspective, but it seems bizarre to me anyway. Even if the target group really is immoral or malicious, they are definitely not subhuman, so this belief is clearly inaccurate by any more conventional definition of accuracy. Functional perspectives’ emphasis on perceivers’ goals and happiness, for me, leads to a far too subjective definition of accuracy. Although functional perspectives have a justifiably important place in psychology, they often do not lead to a useful or reasonable consideration of accuracy.

CONSTRUCTIVIST PERSPECTIVES

Social constructivism (aka social constructionism) is a family of perspectives that emphasize the extent to which people’s beliefs, traditions, and practices create or “construct” social reality (e.g., Berger & Luckman, 1966; Holland, 1997). Many social constructivist perspectives

either deny the existence of an actual reality that is independent of social perception, interpersonal interaction, and sociocultural/political processes; do not address the issue at all; or acknowledge the existence of such a reality but downplay its importance relative to other social processes (Danziger, 1997; Gergen, 1985; Hare-Mustin & Maracek, 1988; Holland, 1997). “Constructivism asserts that we do not discover reality, we invent it” (Hare-Mustin & Maracek, 1988, p. 455). Social constructivist perspectives tend to be highly politicized and, at times, seem more concerned with liberating underprivileged or low-status peoples from, to use some favorite constructivist terms (in their view), oppressive, patriarchal, or Euro-centric hegemonic discourses and practices than with understanding basic social and psychological processes. Not that these are completely mutually exclusive—power and status relations are, in my view, *one* set of important social phenomena, but I would argue that there are many others, too. Constructivists rarely concern themselves with accuracy, except to suggest that accuracy is either impossible or meaningless. The whole notion of assessing correspondence between belief and criterion is likely to be dismissed as a futile search for objectivity because, to most social constructivists, there is no such thing as objectivity.

However, as far as I can tell, constructivist perspectives would have to acknowledge a serious internal contradiction, should they ever consider the accuracy issue at all. On the one hand, accuracy is viewed as meaningless or of trivial importance because all human phenomena are socially constructed. On the other, however, even if human phenomena are entirely socially constructed, this perspective would seem to imply that *there is* a reality that is independent of social perception. Once “we” have created it, it is there, isn’t it? And if it is there, then, presumably, it could be perceived, perhaps even by people not involved in its construction.

Implicit in the very existence of social constructivist arguments intending to expose implicit ideologies, hidden agendas, and social discourses that reify existing power and status arrangements would seem to be the idea that the social constructivists themselves are able to *accurately* identify such ideologies, agendas, and discourses. Exposing a hidden ideology that is not really there would seem to be a pretty silly exercise. If social constructivists can accurately identify such social phenomena, accuracy would appear to be allowed in via the back door. That is, if social constructivists’ perceptions can be accurate, perhaps other people’s perceptions can be accurate, too. Not subscribing to such a constructivist position myself, in large part both because of its heavy-handed politicization and because of this type of internal contradiction, I will leave it to the constructivists to try to sort all this out.

Although this strong social constructivist position has had much more influence within other social sciences (sociology, anthropology, women’s studies) and the humanities (English, history, etc.) than within psychology, at least one psychological review of accuracy has presented a politically neutral and generally less extreme constructivist perspective (Kruglanski, 1989). While skirting the issue of whether there really is a social reality out there independent of individual perceivers, Kruglanski has argued that all criteria ultimately boil down to agreement (see also Kenny, 1994).

Was Michael Jordan the best basketball player of the 1990s? As long as most people agree that he was, then he was (unless you are one of the disagreeers). “Whoa,” you say, “it is not just agreement. Look at his extraordinary statistics. And those six championships in 8 years.” Whether a shot goes in the hoop is purely objective, isn’t it? I think that Kruglanski’s (1989) perspective would suggest not. The shot counts only if the referees agree that it went in. If they agree that it did not go in, then it is not two points.

And this argument becomes progressively more powerful the more fuzzy the attributes or behavior being judged. Is Romain smart? We could give him an IQ test to find out. Of course, if you do not agree that IQ tests are indicative of intelligence, then you would not likely agree with any conclusion emerging from such a test. Is Anushka extroverted? We could find out by observing her behavior in my class, at the next party, and at a dinner with friends. Of course, we could only establish the accuracy of our assessment of her extroversion by examining the extent to which our perceptions and interpretations of her behavior concurred with one another's.

Although I understand this argument, I do not buy it in its absolutist form (*all* accuracy comes down to agreement). In the Michael Jordan case, establishing the rules of the game requires agreement. Once established, though, whether or not that ball goes through the hoop is, in my opinion, a purely objective fact independent of social perception. It may even legally be considered a score by the refs, and if so, it counts, but the refs could have made a bad call. That is, they could be wrong.

In general, and even including relatively fuzzier attributes such as intelligence or extroversion, I think it is often not too difficult to first define exactly what we mean by the construct and then obtain evidence regarding how much of it someone has. Agreement is one very valuable source of such evidence, but it is not the only such source. Just what those sources might be is the central focus of the remainder of this chapter.

Criteria for Establishing Accuracy

TRUTH WITH A SMALL *t*

Criteria and Construct Validity

Identifying criteria for establishing accuracy is an important issue, and sometimes a complex or tricky one, but I see it as no more problematic than establishing the validity of virtually any other social science phenomena. My own general approach to accuracy has been exquisitely articulated by Funder (1995), who likened establishing accuracy to establishing construct validity (Cook & Campbell, 1979). Psychologists and other social scientists spend a great deal of time studying intangible, hypothetical constructs, such as self-esteem, attitudes, mental modules, personality, expectations, schemas, intelligence, stereotypes, prejudice, etc. In general, the myriad papers on these and similar hypothetical, intangible constructs are written as if the authors believe these constructs are real.

I completely concur with Funder (1995, p. 656) that "... although truth exists, there is no sure pathway to it. There is only a wide variety of alternative pathways, each of which is extremely unsure." Please do not misinterpret this as meaning I have all but capitulated to the camp of accuracy naysayers. Although only the gods may know Truth with a capital *T*, the rest of us can establish truth with a small *t*.

What Is Truth with a Small *t*?

The duck test. Truth with a small *t* is evidence, preferably from a variety of sources, indicating that some belief is valid (true, warranted, justified, etc.). In short, the solution to the problem of partially fuzzy, intangible constructs (at least compared to, e.g., distance, heat, etc.) is to

establish their validity through multiple methods and approaches (e.g., Campbell & Stanley, 1963; Cook & Campbell, 1979; Cronbach, 1955). Truth with a small *t* is typically established via rigorous methodological variations of the “duck” test (Block, 1993; Funder, 1995): If it looks, walks, acts, and sounds like a duck, although the possibility remains open that it really is an antelope, sport utility vehicle, alien visitation, or bacterial growth, it is most likely a duck. I am, of course, more likely to confuse a duck with, say, a goose or a gull than with, say, the Statue of Liberty or a praying mantis. Even if I do believe that goose is a duck, however, I further assume that it would, eventually, be possible for me to receive information that corrects my faulty view.

Another peculiar irony: How can researchers study fuzzy, intangible psychological characteristics and assume that no good criteria exist for establishing accuracy? Nearly all published articles in psychology (except for those focusing exclusively on biological or chemical processes) and, indeed, most of the social sciences involve fuzzy, intangible, not directly observable, underlying constructs. Presumably, social scientists only study phenomena that they believe really exist. Who would study stereotypes, schemas, heuristics, achievement motivation, hostility, extroversion, attitudes, self-esteem, or depression if they did not believe such phenomena really exist?

Implicitly, therefore, all psychologists using such constructs, *even those expressing explicit qualms about the viability of accuracy research* (e.g., Fiske, 1998; Jones, 1985; Stangor, 1995), would seem to be logically compelled to accept the idea that the characteristics of people that they study are real. If so, then they are equally logically compelled to accept the idea that there are good criteria for establishing the accuracy of social beliefs, because they would be *the very same criteria* that psychologists use to establish the reality of the constructs they study (this point is reminiscent of Kelly’s [1955] perspective on the *lack* of any fundamental difference between psychologists and laypeople—see the long quote at the beginning of Chapter 10).

How to do this is the point of this section on criteria. I distinguish between four broad classes of criteria: objective criteria, behavior, agreement with others, and agreement with targets. Criteria are objective when that which is being judged is assessed in a standardized manner that is independent of the perceiver’s judgment. Behavior refers to actions of the target(s). Agreement with other perceivers refers to correspondence between the perceivers’ judgments and those of other people. Agreement with targets refers to correspondence between perceivers’ judgments regarding targets and targets’ self-descriptions. I next discuss each of these classes of criteria in detail.

OBJECTIVE CRITERIA

The Simplest Case for Assessing Accuracy: One-Shot Hit or Miss

Is anything ever independent of somebody’s social judgment? I think the answer is a clear “yes.” Consider the following example. During the late 1990s and early 2000s, Mike Piazza was the Mets’ superstar catcher. Let’s say he comes to the plate with the bases loaded and the Mets behind 8–5. I say, “I bet he’s gonna hit a grand slam.” There are only two possibilities: He does, and I am right, or he does not, and I am wrong. There is nothing the least bit difficult or “problematic” about this. The criterion is obvious and objective. Although the rules of baseball can only be established through agreement, once established, the criteria for hits,

homeruns, walks, strikeouts, etc., are mostly independent of human judgment. The role of umpires is primarily to exercise subjective judgment for (the relatively few) close calls, to prevent unruly or aggressive behavior, and to enforce the more esoteric rules of the game—rules that even the players sometimes forget about.

This homerun example is very simple, although perhaps not quite as nakedly simple as this example appears. First, there is the issue of near misses. Maybe I am “wrong” in that he hits the ball off the very top of the fence, but it takes a weird bounce back into the park. Instead of a grand slam, he has hit a bases-clearing double. Although I was, strictly speaking, wrong, I would feel as though my prediction was awfully close, and a helluva lot more accurate than had I predicted, say, that he would strike out or hit into a double play.

Second, there is always the possibility of ambiguity. Perhaps Mike hit a bullet into the seats right down the foul line, which the ump, who is obviously blind, called foul (although with the recent advent of televised replay, such errors have been greatly reduced). Or perhaps that double did not hit off the fence—I saw it hit off a fan in the seats and then pop back in. It “really” (to me) was a homerun.

Although such situations do indeed occur, they are relatively rare: 99 times out of 100, or more, there will be no controversy on whether the ball was a homerun or not. In the event that Mike hits a clear and obvious homerun, we all agree that he had a homerun, but the criteria for evaluating whether or not it was a homerun is not our agreement. We agree because we all saw it clear the centerfield fence. Agreement is a result of accuracy, not a criterion for establishing accuracy.

This type of thing comes up all the time and, I suspect, captures what most people think of when they think about accuracy. Did you predict that Gore would win the 2000 presidential election? If so, you were wrong. You were wrong, even if you believe Gore received the most votes in Florida and *should* have won the election. Unless you forgot all about the Electoral College and what you really meant was that Gore would win the popular vote—if that is really what you meant, then you were right.¹

Did you predict that the American stock market would decline in 2000? If so, you were right. If you did not make this prediction and kept all of your investments in stocks in 2000, no matter how strong your belief that the stock market would go up and was a good investment, you ended up with less money at the end of 2000 than you had at the start of 2000.

Both my election and stock market examples are one-shot predictions with an objective criterion. As with baseball, social processes and agreements are necessary to arrange for political processes and economic investments. But once those social arrangements are in place, all sorts of outcomes occur independent of individual perceivers’ beliefs, predictions, or expectations. These come up all the time in daily life, and the criterion issue is rarely a “problem.”

Many interesting and important social phenomena involve simple, clear, objective criteria. Thus far in this book, I have referred to having a child who plays soccer, so you should know that I am a parent. But take a guess: How many kids do I have? . . . The answer is three. Am I married, divorced, remarried? Married only once and still married (at least as of the time of this writing). Did I graduate high school? Yes. College? Yes. Do I have a master’s degree? No (I went to one of the few PhD programs in the country—Michigan—where it was possible to get a PhD without first getting a master’s degree). Am I a member of a political party? No (I am a registered independent). These are not opinions or fuzzy constructs. They are simple, clear, and objective—not the least bit “problematic.”

Unfortunately, however, finding out whether a single person makes an accurate or inaccurate single prediction is not usually particularly psychologically interesting. We come away with no broad and generalizable principles, and such predictions rarely have much bearing on any psychological theory or hypothesis. Psychologists are more likely to be interested in evaluating how accurate people are at making a particular type of judgment, conditions helping or hindering accuracy, or the processes by which they arrive at accurate versus inaccurate judgments. Usually, therefore, research on accuracy requires investigating many people making one or more judgments or predictions in order to evaluate their accuracy. Such situations, although a bit more complicated than one-shot accuracy with objective criteria, rely on the same, fundamental ideas and, therefore, are not particularly problematic either.

Less Simple, but Still Straightforward: Overall Levels of Accuracy

With objective criteria, it is fairly simple to assess overall levels of accuracy across a wide variety of people and a wide variety of judgments. For example, Archer and Akert (1977) developed a test of nonverbal sensitivity that involved assessing people's ability to accurately identify objective aspects of other people, their experiences, or their relationships. This test included the following situations: People were exposed to two women playing with a baby and talking, and two men who had just finished a one-on-one basketball game. Participants' task was to correctly identify the mother of the baby and the winner of the game. In the actual test, there is a series of such vignettes, and the more questions answered correctly, the more accurately people are at judging others on this test. This test was used to assess people's acuity at judging others primarily on the basis of nonverbal cues (none of the targets ever gave away critical information). This is simple and objective; there is no "criterion problem" here. Indeed, there is nothing methodologically or scientifically difficult at all about assessing accuracy here.

The criterion issue is also a nonproblem in much of the deception-detection literature. That is, one question psychologists have asked is, "How good are people at detecting when others lie?" For example, Ekman and Friesen (1969, 1974) had targets view either a pleasant film or a horrible gory one. All had to inform perceivers that they had just seen a lovely, pleasant film. The research question was how often people can tell when others are liars. Again, there is no criterion problem.

In studies of some aspects of stereotypes, one can also compare people's perceptions to objective data. In one classic study, McCauley and Stitt (1978) compared people's beliefs about differences between African Americans and other Americans to U.S. Census data. Are adult African Americans more or less likely to have completed high school, be on welfare, come from a single-parent home, and so on? Again, the criteria are not "problematic."

With this type of data, one can establish overall levels of accuracy, error, and bias. Do people err on the side of believing what others say? Then Ekman and Friesen's (1969, 1974) studies should have yielded results showing that people underestimate deception (it did, but their research also typically showed that people do better than chance at accurately detecting deception). Do people exaggerate differences between demographic groups? Then McCauley and Stitt (1978) should have found people overestimating the differences between African Americans and other Americans (they did not—people tended to be accurate and, when wrong, to underestimate real differences). These are important and interesting accuracy questions. And the availability of clear, objective criteria in these cases is not "problematic" at all.

Independent and Standardized but Not Universally Persuasive Objective Criteria

Not all people may agree that certain objective criteria are good ones. Such agreement might be irrelevant regarding, say, guessing my number of children, but they become much more relevant when estimating, say, my extroversion or intelligence via a personality questionnaire or standardized IQ test. Is the personality questionnaire a good one? Is it reliable? Valid? IQ tests, in particular, have a long and controversial history (e.g., Gould, 1978; Herrnstein & Murray, 1994; Neisser, et al., 1996).

To the extent that some people do not find such tests credible, they are likely to discredit or dismiss research on accuracy using such criteria. Thus, use of objective but controversial criteria can be viewed as boiling down to agreement (if you agree with the criteria, the study assesses accuracy; if you do not agree with the criteria, it does not—see Kruglanski, 1989). And socially and politically, this is probably how things work. People who do not accept your criteria most likely will not accept your conclusions (whether on accuracy or any other social science topic).

Often, however, what may happen is the reverse: People who do not like your conclusions will come up with arguments against the appropriateness of using your criteria. Chapter 10 has already suggested that this may help explain why social psychologists were much more critical of the criteria used in accuracy research than in self-fulfilling prophecy research, even when the criteria were identical.

Another double standard: The case of cognitive ability tests. The double standard of heavy criticism of criteria for accuracy but acceptance of the very same criteria when used to demonstrate phenomena seeming to provide insights into sources of inequality is not restricted to interpersonal expectations. The same pattern can be observed regarding research on stereotype threat. Stereotype threat was originally the idea that cultural stereotypes about intelligence (e.g., for African Americans) or achievement in some domains (e.g., math for women) leads to anxiety or concern among members of those groups about confirming those stereotypes (it has since been expanded to include all sorts of concerns by all sorts of groups about confirming all sorts of stereotypes). Such anxiety then undermines their academic achievement (e.g., Steele, 1997).

Stereotype threat has gained widespread visibility and acceptance among social scientists (Crocker, Major, & Steele, 1998). Although I think such acceptance is well-deserved (numerous studies have been performed documenting various aspects of the phenomenon), I also find one aspect of such acceptance peculiarly ironic. Despite the frequent objections to IQ tests that periodically appear in various social science, editorial, and intellectual outlets (see, e.g., Jacoby & Glauberman, 1995), I am aware of no social scientific criticism of the use of cognitive ability tests as criteria for establishing stereotype threat-related phenomena. If cognitive ability tests are invalid, then research identifying conditions under which some people score higher or lower on an invalid, meaningless test would not seem to be particularly informative.

Why, then, have cognitive ability tests been the target of so much criticism as measures of intelligence or achievement, but not as criteria with which to establish stereotype threat? The social problems/injustice orientation of many social scientists would likely lead them to be far more accepting of stereotype threat phenomena (which suggests that demographic group differences in cognitive ability test scores result from oppressive cultural stereotypes)

than of research that merely documents the existence of group differences on cognitive ability tests (which is often [mis]interpreted as reifying and essentializing group differences²). For some people, apparently, IQ and other cognitive ability tests are objectionable primarily when they lead to objectionable conclusions.

Politics aside, however, such dismissal is not justified for several theoretical and methodological reasons. Such tests are still *objective* in the sense that they are *independent* of social perception (your score on the personality questionnaire or IQ test is your score, not someone else's perception of your score, and the standards for what constitutes high and low performance are set in advance). Thus, participants' scores or responses on such tests, with all their limitations, are at least untainted by the researchers' or other perceivers' own subjective or interpretive biases.

If I believe that teachers mistreat children from lower socioeconomic backgrounds, this may color my interpretation of a teacher's interactions with her kindergarten class (see my critique of Rist, 1970, in Chapter 6), so that I conclude that she is far more biased than she really is. In contrast, if my "observations" suggest that her social class stereotypes are powerfully self-fulfilling, it would seem to predict at least some social class-based differences in performance on an IQ test by the end of the year. The absence of such differences would (or at least should) make it very hard for me to maintain my position that such stereotypes are powerfully self-fulfilling. Similarly, if the teacher generally believed in large differences in intelligence between children from middle class versus lower social class backgrounds, an absence of social class differences on an IQ test would strongly suggest she was wrong.

A second reason that such attempts to dismiss the validity of research using imperfect or controversial criteria themselves may sometimes be inappropriate is that there is usually some, and sometimes a great deal of, evidence supporting the validity of such tests. Smart people can almost always engage in the intellectual contortions necessary to maintain a cherished viewpoint (e.g., "test x is no good"). Extraordinary contortions, however, are necessary to maintain this conclusion regarding cognitive ability tests, such as most IQ tests, the SATs, and the GREs. Such tests predict (1) academic performance at every level of schooling, from elementary school through graduate school; (2) occupational status and income as an adult, even when the tests are administered to children; (3) job performance; and (4) all sorts of social outcomes, such as the likelihood of becoming a criminal, welfare recipient, or unwed mother (Gottfredson, 1997; Herrnstein & Murray, 1994; Kuncel, Hezlett, & Ones, 2001; Neisser, et al., 1996; Schmidt & Hunter, 1998).

Furthermore, although there are indeed often demographic group differences in scores on cognitive ability tests themselves, in terms of the outcomes such tests are supposed to predict, there is no evidence that modern tests are systematically biased against particular minority, linguistic, or cultural groups (Kuncel et al., 2001; Sackett & Wilk, 1994). In contrast to popular cultural mythology, cognitive ability tests are more likely to be biased in *favor* of members of stigmatized or marginalized groups than against such groups, if the criterion for establishing bias is prediction of performance in school or on the job. On average, members of groups against whom the tests are supposedly biased typically perform no better, and sometimes perform worse, in school or on the job than do people with identical scores from groups against whom the test supposedly favors (Gottfredson, 1997; Kuncel et al., 2001; Sackett & Wilk, 1994). Such conclusions are not restricted to researchers who have been accused of racism; they include numerous researchers who have received the American

Psychological Association's Award for Distinguished Scientific Contribution to Psychology, such as Scarr, Schmidt, Hunter, and Meehl (Gottfredson, 1997).

Cognitive ability tests, such as most IQ tests, the GREs, and the SATs, are among the most highly validated tests within psychology and the social sciences. Those maintaining that such tests are, in some way, "bad" (invalid, biased, etc.) implicitly undermine the credibility of virtually all social science research, because if the best is not good enough, then none of our established measuring instruments (e.g., Robinson, Shaver, & Wrightsman, 1991) and certainly none of the measuring instruments researchers routinely develop "on the fly" (i.e., when no preexisting measure exists, researchers often create a few questions to assess some construct) would be good enough either.

Consequences of the use of imperfect criteria. Similar intellectual contortions would also be necessary to maintain resistance to accepting other criteria. For example, consider the argument that "accuracy in judgments of extrovertedness cannot be assessed" in the face of evidence showing, for example, that perceivers' judgments of targets' extrovertedness predict targets' scores on a questionnaire intended to measure extrovertedness. As far as I can tell, "the scale is bad" would be a far better explanation for a *lack of correspondence* between judgment and questionnaire than for correspondence. That is, if people are right and the scale is wrong, the two will not correlate highly with one another. But if the two do correlate highly with one another, it would appear that what perceivers see in targets corresponds reasonably well with what the scale is supposed to measure.

Thus, use of imperfect criteria will generally lead us to *underestimate* people's accuracy. If so, in sharp contrast to attempts to dismiss the viability of accuracy research because of imperfect criteria, this means that people are probably *even more accurate* than indicated by the evidence emerging from any particular study using imperfect criteria.

For the statistically inclined, this is obviously the case when using a criterion measure with less than perfect reliability. Unreliability in measurement artificially *lowers* correlations between judgments and criteria, so that the greater the unreliability, the *higher* the *actual* accuracy. Concretely, consider a case in which people's perceptions, judgments, etc., of others' extroversion correlate .3 with a self-report extroversion scale. If there is less than perfect reliability in measurement of both perception and criterion, then the true correlation will be higher than .3. If, for example, the reliability of both scales was .9, the true correlation between perception and criterion would be .33; if the reliability of both scales was .75, the true correlation between perception and criterion would be .4 (Carmines & Zeller, 1979).

This would often be equally true with a scale of imperfect or partial validity. Consider a scale that only captures a partial aspect of some attribute. People's judgments, perceptions, etc., if they reflect more aspects of the criterion than the scale being used, may be *more* accurate than indicated by the correlation between judgment and criterion.

This is obvious in the case of objective criteria. Consider an assessment of people's accuracy in judging size. People are simply asked to rate how big a target is. The criteria, however, is height. Height, of course, is only one aspect of size (the others being width and depth). If people use height, width, and depth to judge size, the correlation of their judgments with the (imperfect, partial) criterion of height will be too low. That means their accuracy could be *even higher* than indicated by that criterion.

Similarly, consider the use of an IQ test as a criterion for evaluating the accuracy of judgments of intelligence. If a particular IQ test primarily taps verbal intelligence and people's

judgments include, in addition to estimations of verbal intelligence, creativity, social skill, political savvy, wit, and common sense, then, *if people's judgments closely correspond to actual intelligence defined in this multifaceted way*, the correlation between judgments and criterion will be artificially low. In other words, people would actually be more accurate than indicated by the correlation (correspondence) between judgment and criterion.

Of course, just because such mismatches may lower the observed correspondence between perception and criterion, one cannot infer high accuracy from low correlations. But the argument that any particular criterion in any particular study is "bad" (unreliable, low validity) means that whatever evidence on accuracy is obtained probably often represents a *lower bound* on accuracy. This, in turn, means that people may actually be more accurate than indicated by the empirical evidence of accuracy obtained in that study.

I am not arguing for use of bad criteria. I am arguing, however, that just as imperfect criteria do not preclude the possibility of objective, social scientific research on all sorts of topics (aggression, achievement, identity, social class, etc.), imperfect criteria do not preclude the possibility of accuracy research either.

Independent, objective, but controversial criteria need not be restricted to psychological tests. For example, in many states, one's car insurance premiums increase when one receives a moving violation. One method of adjusting rates involves assigning different violations different amounts of penalty points. More serious violations incur more penalty points. With one of my former insurance carriers, car insurance typically increased by 10% per point, and this premium lasted for 3 years. Insurers were, in essence, claiming that each point worth of ticket predicts, on average, a 10% increase in the likelihood of having an accident for which they have to pay, which then disappears after 3 years. Is this true? I do not know.

I receive a moving violation about once every 5 years, and I have had one accident in 30 years. Lord knows, I would love to dispute the addition of an insurance surcharge on the basis of moving violations. Nonetheless, the criteria (license penalty points) are independent (of the insurance company's judgment) and objective (the police are almost always right when they catch people in moving violations, even me), even if it is not clear that it is a valid predictor of accidents.

The bottom line here is that some types of criteria are subject to debate and controversy. This, however, does not invalidate the research using such criteria. Indeed, even I would have to agree that a driver with no accidents and *no* tickets in 30 years is probably a better driver than one with one accident and a few tickets in 30 years.

In general, the more highly a test or criterion has been validated, the more confidence we can have in its use in evaluating people's accuracy. But there is no hard obstacle to the use of imperfect criteria in accuracy research. The extent to which nearly all measures of social science constructs have been validated or well-established has limitations, thereby inflicting imperfections on *any* research in psychology or any other social science using *any* measure of *any* construct (i.e., all research). Researchers should almost always carefully evaluate the flaws and limitations of their research, including their use of measures. But this issue is no more a "problem" for accuracy research than for any other type of research. Those who claim criteria are a unique or special problem for accuracy research are, at best, implicitly claiming all social science research has the exact same degree of "problem." At worst, they are hypocrites applying a stark double standard wherein the criteria used in research of which they approve is "acceptable" but the (exact same) criteria used in research of which they disapprove is "a problem."

BEHAVIOR

What Is Behavior?

In the broadest sense, anything a person does is behavior. In this sense, answers on a standardized test constitute behavior. Similarly, agreement among judges usually constitutes agreement concerning either targets' behavior or inferences about their attributes based on their behavior (Kruglanski, 1989). Even completely unobservable thinking, feeling, or daydreaming can be considered behavior. Indeed, transmission or inhibition of neurotransmitters can be considered behavior at the level of the individual cell.

None of this is what I mean by behavior. For the social perception contexts addressed in this book, I consider behavior to be observable action, not unobservable thought processes, cell electrochemical transmissions, or underlying attributes. Counting how frequently Natasha smiled is related to but different from determining whether she is thinking happy thoughts or how sociable she is. Smiles are visible. Similarly, measuring how much time Juan spends on some task is related to but different from evaluating Juan's motivation. Again, measuring time on task is a lot easier because it is directly observable behavior, not an underlying attribute.

Along those same lines, "answers on a standardized test" is not what I, or I suspect most social scientists, mean when we use the term "behavior." Such tests are generally designed to measure some attribute that has relevance outside or beyond the testing situation. Intelligence should predict school outcomes, work outcomes, and many life outcomes. Scores on a personality scale measuring hostility should predict levels of hostility, anger, aggression, etc., in all sorts of interpersonal contexts. Although standardized test scores could be considered one class of behavior, it is both possible and useful to distinguish between standardized tests and overt actions.

Does Behavior Boil Down to Agreement?

Some would argue that it does (e.g., Kruglanski, 1989), and for good reasons. In many research contexts, human behavior is measured by having independent judges observe and code the behavior. But there is a difference between the need for people to *observe and record* behavior from the need for people to *interpret* behavior. Behavior itself is observable; personality, beliefs, attitudes, motivation, competence, intelligence, etc., may be inferred from behavior, but they are not directly observable. Observers recording nonverbal behaviors (e.g., smiling, speech disfluencies, etc.), time spent on some task (e.g., proportion of time spent talking in a dyadic conversation, time spent trying to solve an impossible anagram before giving up, time it takes to walk down a hallway, etc.), task choice (how many advanced math courses a person takes, whether a person chooses to go to college, number of parties attended, etc.) are all recording objective aspects of behavior.

Admittedly, sometimes, around the fringes, there may be room for interpretation. Was Fred smiling, or was the corner of his mouth just twitching a bit? Was Elmira staring at Darrin or just daydreaming in his general direction? And it is very useful to have multiple observers record behavior in order to resolve these fringe or ambiguous situations. When that happens, judgments of behavior do boil down to agreement. But, when *behavior* rather than underlying attribute is being recorded, such fringe or ambiguous situations are generally

likely to be the exception, rather than the rule. Thus, there is an important distinction between using observers to record what a person does and using observers to interpret and evaluate the meaning of what a person does. Behavior is what a person does.

AGREEMENT WITH OTHER PERCEIVERS

I suspect that agreement, at first glance, appears to be a poor criterion for accuracy. You and I can both agree on something and both of us could still be wrong. This is obviously true. Agreeing that the world is flat does not make it flat. Thus, agreement, like other criteria, is almost always imperfect.

But agreement cannot be so readily dismissed. At its simplest, if both of us are accurate, we must also agree. If Piazza really did hit a grand slam, and we both saw it, we will both agree that he hit a grand slam. If Hanna is a brilliant student, and we both recognize her brilliance, we will both agree that she is brilliant. And so on. Accuracy requires agreement, although agreement does not *necessarily* imply accuracy. Furthermore, if we disagree, at least one of us must be wrong (at least if we are talking about the same thing).³

This analysis indicates that accuracy must increase agreement. In probabilistic or correlational terms, as accuracy increases, so should agreement most of the time. In fact, this is one of the key ideas in understanding why agreement is often a *good*, if imperfect, index of accuracy. Correlations work two ways. Let's say that A is positively correlated with B. If so, it is equally appropriate to claim that as A goes up, so does B, and that as B goes up, so does A. If accuracy and agreement must be positively correlated, then we also know that, in general or on average, as agreement goes up, so does accuracy. Thus, agreement is an imperfect but probabilistic indicator of accuracy.

In fact, however, even "agreement" itself is a multifaceted construct. Agreement with whom? Experts? Other people? The targets themselves? In fact, all of these types of agreement may, under different circumstances, constitute better or worse criteria for accuracy. The next section, therefore, (1) describes and reviews the types of agreement criteria most commonly used in accuracy research and (2) identifies the major potential limitations of each.

Agreement with Experts

Perceivers' judgments or expectations can be compared to those of experts. For example, personality or psychopathology judgments could be compared to those of professional experts, such as clinical psychologists or psychiatrists; ratings of students' aggressiveness or academic motivation might be compared to those of the students' teachers. For some types of research, nonprofessionals could also be considered "experts" in the sense that they have some sort of unique access to knowledge about the target. Thus, spouses, close friends, and co-workers could be used as experts with whom to compare perceivers' judgments of targets.

Agreement with Experts' Models

Sometimes, experts, such as statisticians, decision-making theorists, or psychologists have developed formal models for what constitutes the most appropriate way to arrive at a judgment. Accounting for regression to the mean, appropriate use of base rates, and use of

consensus, consistency, and distinctiveness in arriving at attributions all constitute formal models of the appropriateness of social judgment (see, e.g., Dawes, 1979; Kahneman, Slovic, & Tversky, 1982; Kelley, 1967; Nisbett & Ross, 1980). Thus, judgments or decisions based on such models may be used as criteria against which to evaluate the accuracy of laypeople's judgments or decisions.

Agreement with Independent Judges

Perceivers' judgments may be assessed after interacting with a target. If the interaction is recorded, for example, by videotape, transcription, etc., people other than the perceiver can be asked to evaluate the target. I call these people "independent" judges because they were not involved in the interaction between perceiver and target; thus, they are independent of the perceiver. Often in such situations, these independent judges may only be exposed to the targets' responses, to further minimize any effects of the perceiver's verbal or nonverbal behavior on the independent judges (e.g., Goldman & Lewis, 1977; Word, Zanna, & Cooper, 1974).

Agreement with Nonindependent Judges

Sometimes, judges may indeed interact with the target, either simultaneously while interacting with the perceiver or in other contexts. Indeed, Kenny's (1994) social relations model, which addresses numerous aspects of agreement, accuracy, and social perception, requires a round robin design, which means that at least four people must interact with and rate one another. Nonindependent judges may be strangers, acquaintances, or experts.

Limitations to Agreement with Other Perceivers

The perspective of probabilistic realism suggests that nearly all criteria will be imperfect to some degree. Agreement is no exception. The fundamental limitation to use of agreement is, of course, that everyone can be wrong.

Even experts may be wrong. Indeed, sometimes, expert judgments or predictions may be little better than those of laypeople (Cronbach, 1955). Psychiatrists, for example, predicted that hardly any teachers would give 450 volts worth of shock to learners providing wrong answers on a test in Milgram's (1974) studies of obedience (even though half or more often did).

In addition, experts may be subject to their own biases. Happy romantic couples, for example, may see each other in an overly positive, almost idealized way (Murray, Holmes, & Griffin, 1996). The doctors and staff at psychiatric hospitals at least sometimes misinterpret patient boredom or hostility as psychopathology (Rosenhan, 1973). Teachers' expectations sometimes color their interpretations of students' performance (teachers sometimes evaluate high-expectancy students' accomplishments more positively than low-expectancy students' accomplishments, even when those accomplishments are identical—Jussim, 1989; Jussim & Eccles, 1992; Williams, 1976).

Experts' models may be contradictory! I provide two very different examples of contradictory expert model predictions: one involving the stock market and the other involving stereotypes. Efficient markets theory (a common economic view of the stock market among academics) states, in essence, that, because whatever is known is already factored into stock

prices, it is impossible to consistently obtain returns that beat the overall market (Malkiel, 1973). The stock market, according to this view, is essentially a random walk with an upward overall trajectory. The practical implication is that one should just buy the market and hold through (unpredictable) ups and downs.

Because of the random walk aspect, however, when one buys low, the market is just as likely (as usual) to go even lower as to go higher, and when one sells high, it is just as likely (as usual) to go higher as it is to go lower. Trying to buy low and sell high, according to efficient markets theory, detracts from one's overall return on investments compared to a simple buy-and-hold strategy. This is because by buying and holding one never sells too low or buys too high (and because repeatedly trying to buy low and sell high incurs more frequent and, therefore, higher commission costs, which reduce one's overall return). One implication, therefore, is that, because the market is a good investment over the long term, one should just buy and hold. One will only get oneself into trouble (reduce one's overall return) if one tries to "time" the market (buy low, sell high) and beat the averages.⁴

Efficient markets theory, however, can be viewed as clashing with one of the main claims of statistical experts. The statistical idea of regression to the average means that extreme values are likely to return over time to the overall mean. This means that if stock valuations are unusually high, one is more likely to lose money (going forward) than make money, and if they are unusually low, one is more likely to make money (going forward) than lose money. If one enters the market when valuations are unusually low, therefore, and exits when they are unusually high, one should be able to beat the overall market (by receiving the unusually large gains that follow low valuations and avoiding the severe losses following high valuations).

Which is true? Don't ask me. I leave this to the stat types to duke it out with the economists.⁵

A similar contradiction occurs in expert models regarding use of stereotypes. Nearly all social psychological theories and perspectives, and much of current cultural mythology, state or implicitly assume that people are acting rationally and appropriately only when they judge others solely and entirely on the basis of those others' personal characteristics, rather than their group memberships. This view argues or implies that people act in a biased, prejudiced, or irrational manner when they allow their stereotypes to influence their judgments of individuals (e.g., Brewer, 1988; Darley & Fazio, 1980; Fiske, 1998; Fiske & Neuberg, 1991; Jones, 1986, 1990; Myers, 1999).

According to widely accepted principles of both statistics (e.g., Bayes' theorem, see, e.g., Kahneman & Tversky, 1973) and philosophy of science (e.g., Krueger & Funder, 2004; Meehl, 1990), however, base-rates, prior beliefs, and expectations *should* influence our interpretations of new evidence (see McCauley et al. [1980] for an analysis focusing specifically on stereotypes). An unlikely claim (that object I saw in the sky was an alien spacecraft) requires a much higher standard of evidence than does a likely claim (that object I saw in the sky was a cloud). Similarly, if I am not 100% sure what that object was, I am far more likely to be correct if my generally accurate expectation colors my interpretation than if it does not (i.e., I am much more likely to be right if I figure it was probably a cloud than if I figure it was an alien spacecraft).

Similarly, these principles also suggest that if I am not sure how tall a person is, I will, on average, be right more often if I estimate any particular male to be a few inches taller than any

particular female than if I estimate them to be exactly equal (i.e., use, rather than discard, my sex stereotype regarding height). They also suggest that if I do not know someone's social class background, I will, on average, be right more often if I estimate that any particular Jewish American is wealthier and more highly educated than any particular Native American. And even if I am highly unlikely to die in a fatal car accident under any circumstances, I am, on average, far more likely to do so if I enter a car with a male younger than 25 doing the driving than if I enter a car with a middle-aged female doing the driving.

Regardless, it cannot be better to both use and ignore prior expectations. Thus, the expert models that have emerged from the stereotype literature conflict with those that have emerged from the cognitive judgment and decision-making literature.⁶ (I return to this issue in Chapter 18 when I address whether using or ignoring stereotypes increases or reduces accuracy in person perception). For now I simply point out that the expert models developed in the judgment and decision-making literature emphasizing the value of relying on base-rates when situations are ambiguous conflict with those developed in the stereotype literature emphasizing the supposedly biased and irrational nature of ever relying on a stereotype to judge an individual.

Bias and self-fulfilling prophecy should be ruled out! For agreement with other independent observers (or their models) to be a *good* criterion for accuracy, the researcher needs to use observers' judgments that are unlikely to be biased or self-fulfilling and are likely to correspond well with the attribute or behavior being judged. The confounding of accuracy with self-fulfilling prophecy is most likely to be a problem when *nonindependent* judges are used, although this can be accomplished with sophisticated statistical techniques and models (Jussim, Eccles, & Madon, 1996; Kenny, 1994). Because of the potential limitations to agreement with other judges (or their models) as a criterion for accuracy, researchers need to justify the validity of agreement with *any particular* group of judges (or their models) as constituting a good criterion for evaluating accuracy.

AGREEMENT WITH THE TARGET

Agreement with targets' self-descriptions can and have also been used as a criterion for assessing accuracy. I distinguish between two broad types of targets' self-descriptions. I use the term "self-reports of behavior" to refer to the actions targets say they have engaged in. Examples might be how many glasses of alcohol they consumed yesterday, how much time they spend exercising each week, how often they argue with their spouse, and how much sleep they get each night. In contrast, I use the term "self-perceptions" to include targets' attitudes, beliefs, feelings, and evaluations of themselves, their characteristics, and their accomplishments. Self-perceptions might include self-esteem, self-perceptions regarding personality traits (independent, assertive, extroverted, etc.), feelings, political positions, self-evaluations of academic or athletic performance, etc. Self-perceptions, therefore, generally involve unobservable, underlying attributes of some type, whereas self-reports of behavior involve overt, observable actions.

Agreement with Targets' Self-Reports of Behavior

Targets' self-reports of behavior can be used as a criterion against which to evaluate perceivers' judgments. Social reality typically constrains bias (e.g., Jussim, 1991; Jussim, Harber,

Crawford, Cain, & Cohen, 2005; Kunda, 1990; and this entire book), so that self-reports regarding specific and objective behaviors may be less likely to be tainted by self-serving, defensive, or impression management biases than are self-reports regarding vague or ambiguous attributes.

To get concrete, a fraternity member's response to "How many alcoholic drinks have you had in the last week?" may be less likely to be biased than is his response to questions such as "How often did you get drunk last week?" Although bias may emerge regarding even the most objective of behaviors, there is a lot more room for interpretation in the "drunk" question ("Well, I did drink two sixes of beer, but that was over 2 hours, and I never really got drunk") than in the "drinks" question. Similarly, responses to "How much money did you donate to charity last month?" are more likely to be constrained by reality than are responses to "How generous are you?" Again, bias is always possible, but even a highly biased person is not likely to interpret buying a pizza as donating to charity, although people may vary a great deal on whether they consider sharing the pizza with a friend as a hallmark of great generosity.

Agreement with Targets' Self-Perceptions

Despite their imperfections, targets' own self-perceptions regarding underlying or ambiguous attributes, such as personality characteristics or dispositions, can be used as a criterion (see, e.g., Judd & Park, 1993; Ryan, 1995). As usual with probabilistic realism, the issue is not whether such a criterion is perfect, because no criteria are perfect. The question is whether the specific self-perceptions being used as a criterion are likely to reflect what the target is like.

Indeed, there are theoretical reasons for expecting self-perceptions to be good criteria, at least sometimes. People have access to much more information about some of their experiences, inner states, relationships, etc., than do outsiders. Because bias can exist side-by-side with accuracy (e.g., Jussim, 1991), even when biases taint self-perceptions, they do not necessarily completely eliminate their validity (see, e.g., Funder, 1995, and Chapter 12 in this book). If so, they can be used in the same manner as any imperfect criteria.

Self-perceptions of academic ability, for example, usually correspond highly with indicators of academic achievement, such as grades and standardized test scores (e.g., Eccles & Wigfield, 1985; Felson, 1984; Parsons, Kaczala, & Meece, 1982). Evaluating the validity of other types of self-perceptions, especially those regarding personality attributes, remains an important area for future research (see, e.g., Funder, 1995). Nonetheless, attributes that are more readily observable, such as sociability or extroversion, tend to generate high agreement between self-reports and observers' ratings (e.g., Funder, 1995; Kenny, 1994).

Although, in the abstract, many people may agree and still all be wrong, it seems hard to maintain this argument in the event of strong agreement between self-reports and observer ratings. To do so, would, in effect, be claiming, "Both Bonita herself and most of those who interact with her consider her to be highly outgoing, but I, being a professionally trained expert, know that she is really shy, withdrawn, and introverted." This is hypothetically possible I suppose, but not very likely.

More highly validated self-perceptions are, obviously, better criteria than less well-validated self-perceptions. And a great many self-report scales have been highly validated (e.g., Robinson et al., 1991). There currently is, however, relatively little social science evidence

regarding the validity of many types of self-perceptions. The theoretical reasons suggesting that self-perceptions are often likely to be valid to some degree, however, strongly suggest that the default assumption within the social sciences should be that self-perceptions can generally be used as a criterion, as long as their limitations are also carefully understood and acknowledged. Only when a particular type of self-perception has been empirically *invalidated* to the point of uselessness would it be clear that it should not be used.

Limitations to Agreement with Target Self-Reports and Self-Perceptions

Targets, of course, are imperfect themselves. Although they may have unique access to certain types of information (personal experiences, feelings, etc.), many people are subject to both motivated and unmotivated errors and biases (Kunda, 1990; Nisbett & Ross, 1980). Memory is imperfect and potentially subject to expectancy-related biases (see Chapters 5 and 10), so that self-reports of behaviors may often be imperfect records of behavior.

Of course, researchers have also developed a slew of methods for improving the accuracy of self-reports (e.g., daily diary methods, where people record the events of the day when memory is fresh; or beeper methods, where people will carry around a researcher's beeper and stop and write down what they are doing, feeling, etc., whenever beeped). Although even these types of methods may not completely eliminate error and bias, they have become widely used methods because of their demonstrated validity for many types of self-reports (e.g., Hedges, Jandorf, & Stone, 1985; Räikkönen, Matthews, Flory, Owens, & Gump, 1999).

The problem of motivated biases may often (though not always) be greater when using self-perceptions (rather than self-reports of behavior) as a criterion, in part because of the common tendency for most people to view themselves in self-serving ways (Myers, 1999; Kunda, 1990). In addition, many people may either lie outright or slant their responses in such a manner as to present themselves in as socially desirable, moral, and competent a manner as possible (Paulhus, 1991, 1998). Although this means that the overall *average* level for some self-perception may often be too favorable, this may not prevent self-perceptions from being a good *correlational* criterion for assessing the accuracy of perceivers' beliefs (see Chapter 12 for a detailed discussion of this issue).

To get concrete, Bertha may think she is a great athlete and Nyesha may think she is a good athlete. Both may be overestimates (Bertha may only be pretty good and Nyesha may be pretty average). But if their degree of self-inflation is similar, it may be true that Bertha is more athletic than Nyesha. So a coach who views Bertha as more athletic than Nyesha would be correct (and the coach's views would correlate well with Bertha's and Nyesha's self-perceptions).

Sometimes, however, people vary in their degree of self-inflation. When this seems like a possibility, researchers can assess individual differences in such biases using any of a variety of questionnaires (see, e.g., Paulhus, 1991, 2002). By statistically controlling for this tendency, one can obtain less biased, more valid self-perceptions. Thus, even biased target self-perceptions may, under many circumstances, constitute good, if imperfect, criteria for assessing the accuracy of perceivers. In practice, researchers considering the use of particular self-perceptions as criteria need to thoughtfully consider their limitations and develop a convincing case that, in the context under study, they are likely to be good criteria.

HYBRID CRITERIA

Some criteria may be blends of the types listed previously in this chapter. Is graduating from high school an objective accomplishment, is it behavior, or is it agreement? It is probably a bit of all three, but who cares? It could still be used as a criterion for evaluating, for example, the tendency of teachers' expectations to be accurate or self-fulfilling. If Rosenthal and Jacobson's (1968a) "late bloomers" were more likely to complete high school, it would sure look like a self-fulfilling prophecy, regardless of whether doing so constitutes an objective criteria (once the degree is awarded, it is no longer a matter of opinion), behavior (completing the required coursework), or agreement (school personnel has to agree that the bloomer completed the required coursework). Similarly, if teacher expectations predict without causing high school graduation rates, then those expectations would be accurate, not self-fulfilling.

Similarly, self-report of a behavior could be considered an imperfectly reliable report of behavior rather than another type of agreement. One could make the case both ways, but, again, it does not really matter. Use of self-reported behavior, if well-validated, could often be one useful criterion against which to compare perceivers' judgments.

In contrast to the simple knee-jerk manner in which self-reports can seemingly be dismissed due to their widespread and partially justified reputation for susceptibility to self-serving distortions, highly validated self-report scales may often constitute one of the best forms of hybrid criteria. The key concept here is "highly validated." Highly validated self-report scales, by definition, have been shown to successfully relate in a wide variety of ways to a wide variety of phenomena that they *should* be related to. Questionnaire developers have often gone to great lengths to show that self-report scales assessing social skills, psychopathology, extroversion, motivation, self-esteem, etc., (1) correlate well with other similar measures (e.g., the correlation among measures of self-esteem are usually quite high), (2) predict overt behaviors (e.g., self-described extroverts are noticeably more bubbly in social situations than self-described introverts), (3) correlate well with theoretically related life outcomes (e.g., measures of academic motivation often predict school achievement), and (4) correlate well with others' views of those completing the questionnaire (see, e.g., Robinson et al., 1991).

In such cases, "mere" self-report scales, by virtue of their well-known and established relationships with independent judges and/or with a variety of objective, behavioral, and/or other self-report criteria, have essentially become hybrid criteria of the best kind. In other words, because such scales have been shown to relate to many of the major criteria for establishing accuracy, it is no longer necessary for each accuracy study to reinvent the wheel and obtain new validity evidence using these same criteria. By virtue of using a highly validated self-report scale as a criterion for assessing accuracy, because of its empirically demonstrated relationships with all sorts of other measures, one is indirectly using multiple criteria in one fell swoop.

I am not making the case here that researchers can necessarily assume that all or even most self-report scales constitute excellent hybrid criteria for accuracy. When little evidence bears on the validity of such a scale, all that one has is a scale of unknown validity. Although even such scales can be used as criteria, they are obviously less than optimal and should be interpreted with great caution. But highly validated self-report scales are another animal entirely—rather than being a mere crutch of convenience, they are not so readily dismissed with a derogatory "it's just self-report." By virtue of being well-validated, such scales may often

constitute some of the clearest and best criteria against which to assess the accuracy of social perception.

Thus, whether or not one subscribes to Kruglanski's (1989) view that all criteria boil down to agreement does not really matter. It is useful to understand different types of criteria, because, regardless of whether one considers them qualitatively different or variations on an agreement theme, it is clear that there are differences, for example, between standardized tests, expert models, objective behaviors, and self-reports. The very fact that there are so many potential sources of criteria for evaluating the accuracy of some social perception is a major *strength*, not a weakness, of research on accuracy shall be discussed next.

Why the Inherent Imperfection of Most Criteria Does Not Preclude the Study of Accuracy

Probabilistic realism occasionally provides a gold standard for establishing accuracy. Piazza either did or did not hit a homerun. Except in rare cases, that is not a matter of opinion.

More often, however, such a gold standard is not available. This, however, does not come close to justifying the conclusion that we should just throw in the towel. Establishing that a social belief or perception is accurate is much like establishing validity in social science research. Multiple methods and approaches are generally required for establishing the validity of any construct in the social sciences (e.g., Campbell & Stanley, 1963; Carmines & Zeller, 1979; Cook & Campbell, 1979). In much the same manner, the strongest and clearest evidence regarding accuracy comes from research that typically uses multiple measures and methods to establish the accuracy of social perception (e.g., Cronbach, 1955; Funder, 1987, 1995; Kenny, 1994). This does not mean that research examining accuracy using a single method or criterion is uninformative. Such research, however, is typically less definitive than research using multiple methods or criteria (except when that single criterion itself has previously been validated using multiple criteria!).

I say it is a duck. Does it look like a duck? If so, so far, I am right. At least, I am as right as can be determined from the criterion that we used. I could be wrong, but there is (1) evidence that I am right and (2) no evidence yet that I am wrong. Does it walk like a duck? If not, perhaps I am wrong, but now we have an interesting research question, not some sort of fatal flaw in the entire accuracy assessment enterprise (see also Funder, 1995). What might look like a duck but does not walk like a duck? Maybe it's a ducklike mammal (this is a tangent beyond the scope of this book).

If it does walk like a duck, however, we have even more evidence that I am right. Does it sound like a duck? If so, it is really beginning to seem like I am right, although we will never know this, or anything else, with absolute 100% certainty—just like theory in all the social sciences, all the sciences, and daily life.

Notes

1. While we are on the topic of accuracy, a day or so after election day in 2000, when it seemed that each recount of the vote was giving Bush a smaller and smaller lead, I said to my wife, "Lisa,

it's going to end up with Bush winning by a single vote." I was right, but I did not realize that that vote would be a Supreme Court vote, not a Florida voter's vote.

2. Claims that some test is valid, accurate, or unbiased, when there are average group differences in scores on the test, is sometimes misconstrued as reifying or essentializing group differences. Such an interpretation is ill-founded for several reasons. If valid, all that a test does is appropriately measure what it is supposed to measure. A test, by itself, cannot possibly indicate how or why a person or group scores as they do. Tests themselves are mute regarding whether group differences are permanent and fixed or temporary and malleable. For example, if discrimination takes a disproportionate toll on some people's academic experiences, then those people would likely perform more poorly on cognitive ability tests. Such a state of affairs does not invalidate such tests, however, because they are designed to assess how much intellectual competence, skill, or ability people have, not how they arrived at their level of skill. To use another example, people who have been restricted from athletic activities would likely become poor athletes. That does not mean that assessments of their athletic prowess (speed, strength, agility, etc.) would be biased. In the same way, a person or group whose intellectual or academic opportunities were limited would likely develop weaker intellectual skills. Although this might mean those people were victimized by discrimination, it would not mean that *assessments* of their intellectual acumen are biased.

3. We do need to be careful in considering what constitutes "the same thing." You and I could disagree about how pleasant, extroverted, or intelligent Fred is. If Fred acts differently with you than with me, then we both could be right (Swann, 1984). So if we are talking about Fred in general, both of us may be partially right and partially wrong, because perhaps there is no "Fred in general"—there is only Fred interacting differently with different people.

4. For the financially uninitiated, "buying the market" is an expression—it does not literally mean buying all outstanding stocks. It does mean, however, that one can buy, for example, shares of every stock sold in the United States, or Japan, or Asia or Europe, or South America, etc., by buying a stock index fund. The most common U.S. index funds are those that invest in all 500 companies that make up the S&P 500, which consist of the 500 largest publicly traded companies in the United States (which constitutes about 70% of the value of the stock of all U.S. companies). The performance of such a fund is, of course, nearly identical to the performance of the S&P 500 stock index (it is very slightly less primarily because of fees). Common market averages in the United States are the Dow Jones Industrial Average, the S&P 500, and the NASDAQ.

5. My hunch is that the efficient market types are right much of the time, but not all of the time, so that buy and hold is probably usually, but not always, the best strategy. Exceptions could occur when people act irrationally en masse. This can be seen in occasional market manias and depressions—see the stock market discussion in Chapter 6 of this book and the first three chapters of MacKay's (1841/1932) *Extraordinary Popular Delusions and the Madness of Crowds*. Those chapters in Mackay's book address the irrational stock bubbles produced in the early 18th century in France and England by the excessive enthusiasm for companies actually or sometimes only allegedly involved in the colonization of the New World. It also addresses the insane 17th-century tulip mania in Holland—and all three manias seemed to me to strikingly resemble the technology mania of the late 1990s. Similarly, the greatest bull market of the 20th century (roughly 1982 to 2000), which saw the Dow Jones Industrial Average increase over 10-fold in a mere 20 years, started when stock valuations were unusually low. Thus, although buy and hold may be the best strategy for years at a time, it may handsomely pay off to remember about regression to the mean in the few, rare situations where market valuations reach astronomical pinnacles or plummet into an economic abyss.

6. In my view, the decision-making models are correct and the stereotype models are usually, though not always, incorrect with respect to their (often implicit) assumptions regarding what constitutes accuracy and bias. Although this will be addressed in detail in a later chapter, the main points are that (1) people only hold expectations that they believe are correct and (2) one's judgments will be more accurate if accurate expectations do influence judgments of individuals (in the absence of full and complete knowledge about all relevant aspects of those individuals—i.e., most of the time) than if they do not.

The stereotype models can also be correct, but they require the assumption that the stereotype-based expectation is inaccurate. Judgments that rely on inaccurate expectations, of course, will be less accurate than judgments that ignore inaccurate expectations. I suspect that the clash between the stereotype models and the normative decision-making models results from the often implicit (but rarely tested) assumption held by many stereotype researchers that stereotypes are generally inaccurate. With this assumption, there is no conflict between the two sets of models. The prevalence of this assumption, in the absence of empirical evidence demonstrating universal stereotype inaccuracy and in the presence of evidence demonstrating at least some accuracy in stereotypes (e.g., Judd & Park, 1993; Lee, Jussim, & McCauley, 1995; McCauley, et al., 1980; Swim, 1994), which will be addressed in detail in Chapters 15 through 19, can be viewed as another manifestation of the powerful and pervasive influence of the intellectual zeitgeist, especially in the stereotype literature, emphasizing error and bias.

12 Accuracy

COMPONENTS AND PROCESSES

A STOPPED CLOCK is right twice each day. That does not make it a good clock.

What does this have to do with social perception? More than it seems. Let's return to my successful prediction (from Chapter 11) that Mike Piazza would hit a homerun when he came up to bat with the bases loaded. That makes me look like a pretty darn good baseball prognosticator, doesn't it?

Not necessarily. Maybe *I always* predict that Mike will hit a homerun. Maybe I always predict *everyone* will hit a homerun. And it could even be worse than that: Maybe I always predict all baseball players will do great things, whether in the field, at bat, or on the base paths. (This would be logically absurd, because it would mean that I would predict both that Mike would hit a grand slam and that the pitcher would strike him out. A detailed discussion of people's ability to hold mutually exclusive beliefs [see, e.g., Dawes, 2001], however, is beyond the scope of this book.)

Even though I might have happened to have been right that one time, I could not necessarily be considered a particularly astute judge of baseball. One could think of my prediction regarding Mike's at bat as stemming from several sources or components: (1) my overall tendency to think well of baseball players; (2) my overall tendency to predict that batters will hit homeruns (over and above my general tendency to think well of players); (3) my overall tendency to think that Mike is a good hitter (over and above my general tendency to think well of players); and (4) my specific tendency to predict that Mike will hit a homerun (over and above my tendency to think well of players; to predict that they, in general, will hit homeruns; and to think well of Mike as a hitter in general—OK, I admit it, even I am getting dizzy at this point). Each component of my prediction can be accurate to some degree, and each contributes both to my prediction for Mike and my overall likelihood of being accurate (across lots of judgments or predictions).

Components of Accuracy

This type of thinking inspired Cronbach's (1955) (in?)famous review and at least two other more recent perspectives (Judd & Park, 1993; Kenny, 1994), all of which identified several processes contributing to social perception and which argued that accuracy needs to be separated into different components reflecting these different processes. This section describes each of these three componential approaches to the study of accuracy.

Cronbach's Components

Cronbach's (1955) analysis suggested that each perceiver's judgment consisted of several components: elevation, differential elevation, stereotype accuracy, and differential accuracy. Each component is discussed next.

Elevation accuracy. Do I see other people through rose-colored glasses or am I a nasty cynical malcontent? **Elevation** refers to my general tendency to over- or underestimate people on the attributes being judged. It corresponds, in the baseball example, to my tendency to predict good (or bad) things for all players all the time.

Elevation accuracy addresses whether a perceiver rates targets, overall, more or less favorably than indicated by the criterion. It is the difference between (1) the average of all of a perceiver's ratings of all targets across all judgments and (2) the average of all targets across all criteria (Kenny, 1994). Thus, there is a single elevation accuracy score for each perceiver's judgments regarding targets.

Let's say I was asked to predict several players' (1) likelihood of getting a hit, (2) likelihood of getting a walk, and (3) likelihood of batting in a run with runners in scoring position.¹ Elevation accuracy would be assessed, for example, by comparing my overall probability estimate, averaging over all players and all three judgments, to the actual overall probabilities, averaging over all players and all judgments. For example, Chris predicts the players on the San Francisco Giants to hit .300, get a walk once every 20 at bats (.05), and drive in a run 40% (.40) of the time with runners in scoring position. The average of these averages would be .250. If, on average, the Giants' players actually hit .250, got a walk once every 25 at bats (.04), and drove in runners in scoring position 31% of the time, the overall average of these averages would be .200. Thus, Chris's elevation accuracy score would be 0.05 and would indicate, in this particular case, that he generally expects these players to do better than they actually do.

In a typical social perception case, elevation accuracy would represent the difference between a person's average ratings across several targets on several characteristics (e.g., friendliness, intelligence, conscientiousness) and the actual average of all targets' scores on criteria reflecting those characteristics (less difference, more elevation accuracy). When both the judgment and criterion are on purely subjective scales (e.g., a 1 to 7 scale going from "not at all" to "a great deal"), as in many social perception studies, elevation typically has little relevance to accuracy. Instead, it primarily reflects *response bias*: people's tendency to give responses at higher or lower ends of the scale (e.g., Cronbach, 1955; Kenny, 1994). Furthermore, elevation accuracy cannot be readily assessed when judgment and criterion are on different scales (e.g., if I rate how good the players are on a 1 to 7 scale, elevation cannot be assessed if the criteria are percentages).

Differential elevation accuracy. Let's say you rate Professor Smith as a better teacher than Professor Jones. **Differential elevation** refers to a perceiver's tendency to rate one target higher than another, averaging over all ratings of each target. This can be considered a "target effect" in that it represents mean differences (averaging over all judgments) between targets.

Differential elevation accuracy addresses whether your differential perception of Smith and Jones is correct. It refers to the correspondence between (1) a perceiver's ratings of each of several targets, averaging over all judgments (e.g., if there are two targets, there are two sets of averaged ratings; if there are three targets, there are three sets of averaged ratings; etc.), and (2) each target's actual standing averaging over all criteria (e.g., if there are three criteria, each target's "actual standing" is the average of his or her criterion scores). Differential elevation accuracy indicates how closely a perceiver's ranking of targets on the traits or characteristics (overall) being judged corresponds to the targets' ranking on the criteria (overall).²

Differential elevation accuracy answers the question of whether my perception of Mike Piazza as a better hitter (averaging over all hitting categories) than Jorge Posada is correct. That is, according to hitting criteria (batting average, homeruns, RBIs, etc.), is Mike actually a better hitter than Jorge? In a social perception case, differential elevation accuracy might indicate whether my belief that Alice is more competent (averaging over ratings of intelligence, responsibility, and social skill) than Susan is actually true. That is, averaging over the criteria for intelligence, responsibility, and motivation, is Alice actually more competent than Susan?

Stereotype accuracy. Does Fred believe that high self-esteem is more common than high intelligence among a group of targets? Does the perceiver rate some traits as being more common than others? The **stereotype**, in Cronbach's system, is the tendency to see some traits as more common than others, averaging over all targets. It probably would have been better to call it a "trait effect," because it represents people's perceptions of the prevalence of each of several traits among a group of targets. Thus, if each of three targets is rated on each of five traits, there will be five trait effects, one for each trait.

Of the traits being rated, do people see those that are most and least common as actually being most and least common? **Stereotype accuracy** refers to the perceived versus actual relative prevalence or ranking of the traits, averaged across all targets (as such, it has nothing to do with what most people usually think of as social stereotypes regarding, e.g., race, class, sex, etc. and nothing to do with stereotype accuracy as discussed elsewhere in this book).

In the baseball example, stereotype accuracy might address the question: Does the perceiver realize that the probability of driving in a runner from scoring position is higher than the probability of getting a hit, which, in turn, is higher than the probability of getting a walk? In a social perception case, it might mean recognizing the relative prevalence of friendliness, intelligence, and conscientiousness among the targets being judged (e.g., perhaps all are friendly, some are high in intelligence, and only one is conscientious). Like elevation, however, if judgment and criterion are measured on purely subjective scales, stereotype accuracy scores would most likely primarily reflect response bias (which ends of the scale people tend to use *when judging that particular attribute*), rather than anything substantively related to accuracy.

Differential accuracy. After accounting for (by removing) elevation, differential elevation (target effect), and the stereotype (trait) effect, does the perceiver see Bill as more articulate but less moral than George? If so, this constitutes the perceiver's **differential** in judgments about Bill and George. This can be considered a uniqueness component to social perception,

because it reflects the perceiver's specific judgments about the degree or level of a specific trait found in a particular target, rather than general tendencies to view the target's characteristics or the particular trait as being more or less prevalent (Kenny, 1994). Thus, if there are three perceivers and five traits, there will be 15 uniqueness effects (one for each perceiver \times trait combination).

Differential accuracy represents people's ability to rank order targets on each specific trait. Out of ten players, Piazza might have the second highest batting average; he might have the highest probability of knocking in a run with runners in scoring position; but he might only have the fifth highest probability of getting on base by a walk. How well people's judgments or predictions correspond to this actual ranking would reflect differential accuracy. In a social perception case, differential elevation accuracy would indicate how well a teacher's belief that Louisa is smarter, wilder, and more ambitious than Kendra corresponds with their actual relative amounts of smarts, wildness, and ambition.

Cronbach's components as ANOVA. For the statistically inclined, it may help to point out that Cronbach's components, which appear highly complex and hard to follow in the original 1955 article, can be simplified into a two-way analysis of variance (Kenny, 1994). For the statistically uninitiated, ANOVA is a statistical technique commonly used in psychological experiments for determining how much each of two variables, independently and in combination with one another, predict or explain some outcome. In Cronbach's system, the ANOVA factors are trait and target (for the statistically uninitiated, the constant below is simply the grand mean of all observations; the trait and target effects are deviations from the grand mean):

$$\text{Judgment} = \text{Constant (elevation)} + \text{Target main effect (differential elevation)} + \text{Trait main effect (stereotype)} + \text{Target} \times \text{Trait interaction (differential)}$$

This equation only describes the components of the judgment. To assess accuracy, these components would need to be compared to the *same component score* on the criterion (which would be obtained by replacing "judgment" with "criterion" in the equation; everything else remains the same but would refer to target behavior or trait rather than perceiver judgment). For those of you interested in seeing Cronbach's components broken out into all their gory detail, Kenny (1994, pp. 117–121) provides a relatively clear concrete example involving three perceivers, three targets, and three traits.

Kenny's Social Relations Model

Kenny (e.g., Kenny, 1994; Kenny & Albright, 1987; Kenny & DePaulo, 1993; Kenny & LaVoie, 1984) has been a prolific researcher in the areas of accuracy and agreement, using a componential model that is related to, but different from, that proposed by Cronbach (1955; see, e.g., Kenny, 1994; Kenny & Albright, 1987, for detailed discussions of similarities and differences). Kenny's *social relations model* (SRM) partitions social judgment into four factors: A constant (elevation), a perceiver effect, a target effect, and a perceiver \times target interaction or uniqueness effect (plus error, which refers to random error in measurement of a judgment).

The SRM differs from Cronbach's components in several important ways. First, it is intended to be a broad and general model for assessing many different aspects of social perception, of which accuracy is only one. Second, research using SRM typically focuses on perceptions regarding one trait at a time, rather than the multiplicity of traits addressed by Cronbach's components. Third, however, it also typically focuses on several perceivers, rather than the one perceiver (at a time) that was the focus of Cronbach's analysis. Thus, SRM research might perform one analysis to find out how accurately Dave, Charles, and Bella perceive each other's intelligence and another to find out how accurately they perceive each other's friendliness. Whereas Cronbach partitioned the judgment into target, trait, and target \times trait components, Kenny partitioned judgment into target, perceiver, and target \times perceiver components.

Elevation accuracy. Is there a general tendency for people to see others as better or worse, overall, than they really are? SRM starts with an *elevation* score (constant) that is similar to Cronbach's. It represents all perceivers' average ratings of all targets on the trait. It can be thought of as a grand, overall mean rating, averaging over all perceivers and all targets.

Elevation accuracy refers to the extent to which this average rating corresponds to the average score of all targets on the criterion. Thus, there is only a single elevation accuracy score for any particular group of perceivers and targets. Furthermore, it can only be obtained if the judgment and criterion are measured in the same units (e.g., if both are on a 1 to 7 scale, it can be assessed; if one is on a 1 to 7 scale and the other a 1 to 100 scale, it cannot be assessed).

For example, consider a hypothetical group of little leaguers asked to predict each other's batting averages. Overall (averaging across all predictions for every person in this group), they predict all other kids will bat .290. By the end of the year, however, the kids only bat .270. Therefore, .20 (.290 - .270) would be their elevation accuracy score (i.e., they overestimate their ability to hit).³

Perceiver accuracy. How well does a perceiver's overall ratings of all targets correspond to what those targets (on average) are actually like when interacting with that perceiver? The *perceiver effect*, in SRM, refers to each perceiver's average rating of all targets (after subtracting out the elevation score). *Perceiver accuracy* refers to how well each perceiver's overall ratings (i.e., averaging over all targets) corresponds to the targets' overall average on the criterion when interacting with that perceiver. There will, therefore, be one perceiver accuracy score for each perceiver.

In the little league example, consider Lillian, who is a good pitcher and knows it.⁴ Even though the rest of league bats .270, they only bat .220 against her. She believes, however, that their average against her is .230. Does this mean that she underestimates just how good she is? Not necessarily. Remember that in this example, these little leaguers have a general tendency to predict that the other kids hit better than they really do (see the elevation example). Lillian believes the other kids are .060 worse against her than they are in general (.290 - .230 = .060), which actually underestimates how well the other kids hit against her (on the criterion, .270 - .220 = .050)—the kids only hit .050 worse against her, not .060 worse.

Whether the perceiver accuracy score is conceptually meaningful or a methodological nuisance, however, is often unclear. It could represent a genuine tendency on the part of the perceiver to consistently over- or underestimate targets. This would *seem* to be the case in the little league example, but that is not clear. This is because the perceiver accuracy score may

also reflect response bias—the perceiver may merely tend to use the judgment scale differently than do other perceivers.

For example, Lillian may not believe that, overall, the average player hits .290 or .270. Perhaps she thinks the average player only hits .230. If this was the case, then her overall prediction of .230 would mean that she sees herself as only an average pitcher. Thus, the precise meaning of her prediction that other kids hit .230 against her is unclear. It may mean that she underestimates how well other kids hit against her, but it may also reflect differences between how Lillian and the other kids interpret and use the batting average scale.

Generalized (target) accuracy. How accurately, on average, is a particular target viewed by others? The target effect refers to each target's mean rating averaged over all perceivers. Kenny (1994) refers to the correspondence of this mean rating with the target's overall average on the criterion as *generalized accuracy* (I would prefer a label such as “generalized target accuracy” because it reflects how accurately, overall, a target is viewed by a group of perceivers, but I am following Kenny's terminology here). Thus, there is a generalized accuracy score for each target (e.g., if there are three targets, there are three generalized accuracy scores).

Let's say, on average, Lillian's teammates believe she is a great hitter and predict that her batting average is .400, when, in fact, although she is good, she is not quite that good and her batting average is actually .350. Her teammates overestimate her batting skill by .050. This is almost, but not quite, her generalized accuracy score. That is because we have not yet subtracted out the elevation component. Remember, these kids overrate everyone by .20. If they overrated Lillian by .20, they would simply be viewing the *difference* between her and the other kids as dead-on accurate. Thus, the .20 elevation effect needs to be subtracted out. Lillian's generalized accuracy score would be .30, not .50, which would still mean the kids overestimate her hitting ability (compared to other kids), but not by quite as much as it first seemed.

Dyadic accuracy. How accurately does the perceiver view that target's behavior *with that particular perceiver*? Kenny (1994) refers to differences in how a target actually behaves (reality) with a particular perceiver (as compared to with other perceivers) and differences in how a perceiver judges (perception) a particular target (as compared to other targets) as uniqueness or relationship effects. Such effects reflect unique characteristics of the relationship of this particular perceiver with this particular target. The more closely these two relationship effects (perception and reality) correspond to one another (i.e., the more highly correlated they are), the higher the *dyadic accuracy* (accuracy within that particular pair, or “dyad”). Because the math begins to become laborious when computing this effect (see Kenny, 1994, for a clear and complete concrete example), I present a simplified conceptual example below.

Lillian's team is ahead 5–4, with two out in the bottom of the sixth inning (little league games only go six innings), but the opposing team has runners on second and third. The current pitcher, Lillian's teammate Joe, is obviously tired, and the opposing team's best hitter, George, is coming up to bat. The coach brings in Lillian to pitch to George. Lillian could walk George, which would bring up the other team's second best hitter. However, Lillian remembers that, even though George has already hit two homeruns this game, almost every time in the past when Lillian has pitched to George, she has gotten him to either strike out or pop out by pitching high, slow balls to him. Although George's overall batting average is .500, he has gotten only 1 hit in 10 at bats (.100) against Lillian.³

Lillian never heard of Kenny's SRM and does not go through anything remotely resembling the hairy componential computations required to estimate dyadic accuracy. Nonetheless, her understanding that, although George is generally a very good hitter, he is not very good against her, is, in essence, dyadic accuracy. So she decides not to walk him and pitches a slow, high ball, which George flails at and pops up behind home plate. The catcher makes the catch, the game is over, and Lillian's team wins a tough one—all because of dyadic accuracy.

Kenny's components as ANOVA. For the statistically inclined, it may help to point out that, like Cronbach's components, Kenny's components, which also often appear highly complex, can be simplified into a two-way analysis of variance (Kenny, 1994). Whereas Cronbach's ANOVA factors are trait and target, Kenny's are perceiver and target:

$$\text{Judgment} = \text{Constant (elevation)} + \text{Perceiver main effect (perceiver effect)} + \text{Target main effect (target effect)} + \text{Perceiver} \times \text{Target interaction (relationship effect)}$$

To assess accuracy, each component would be compared to the parallel effect on the criterion (see Kenny, 1994, pp. 129–134, for a detailed presentation).

Judd and Park's Full Accuracy Design for Research on Stereotypes

Judd and Park (1993) developed the first componential model focusing on explaining sources of accuracy and inaccuracy in social stereotypes. They identified four main components of judgments regarding groups: elevation, perceiver group, target group, and attributes. Because Judd and Park's (1993) componential model is not identical to those of Cronbach (1955) and Kenny (1994), I discuss them here. Because they are so similar, however, and because they are even more mathematically complex than Cronbach's or Kenny's approaches, I present only a brief simplification of their main ideas.

Do people, in general, over- or underestimate others' attributes? **Elevation accuracy**, as in the other componential models, is the overall difference between judgment and criterion, averaging over all perceivers, targets, and attributes. Because it involves averaging over *all* attributes, this component does not have much substantive meaning. In a study where people evaluate several groups' intelligence, competitiveness, and social skill, the elevation component merely indicates that adding together all people's ratings of all target groups and all three attributes produces a higher or lower number than obtained when adding together all target groups' scores on all three attributes on the criterion.

Does one group tend to over- or underestimate others more than everyone else? The **perceiver group effect** is an overall tendency for one group of perceivers to over- or underestimate all the attributes (added together) in other groups (i.e., beyond the elevation effect).

Are all the attributes (added together) of one target group consistently over- or underestimated? The **target group effect** is an overall tendency for one group of targets to have all their attributes (added together) over- or underestimated (beyond the elevation effect).

Do people see other groups (in general) in a stereotypical manner? The **attribute effect** represents an overall tendency to over- or underestimate the prevalence of a particular type or class of attributes. When attributes are chosen for each of two groups so that attributes that are stereotypic for one group are counterstereotypic for the other, the attribute effect

becomes a “stereotypicality” effect—the tendency to view groups as more or less stereotypic than they really are. A general tendency to overestimate stereotypical attributes and underestimate counterstereotypical attributes represents a general tendency (across target groups) for the stereotype to exaggerate real differences. A general tendency to underestimate stereotypical attributes and overestimate counterstereotypical attributes represents a general tendency (across target groups) for the stereotype to underestimate real differences.

Like the other componential approaches, Judd and Park’s (1993) full accuracy design was modeled after an analysis of variance—but with three ANOVA factors (perceiver group, target group, attributes, and all two-way and three-way interactions) rather than the two of Cronbach and Kenny’s SRM. Although a full discussion of those factors is beyond the scope of this chapter, the three-way combination is particularly important to the study of stereotype inaccuracy because it tests for in-group bias. The Subject group \times Target group \times Attribute factor tests whether stereotype exaggeration or underestimation (the attribute effect) is more likely to occur when (1) people from Group A judge people from Group B *and* when people from Group B judge people from Group A than when (2) people from Group A judge people from Group A *and* people from Group B judge people from Group B. If, for example, stereotype exaggeration only occurs when people judge groups *other* than their own, one would have an in-group bias effect.

“Must” Components Be Assessed in All Accuracy Research?

Ever since Cronbach’s (1955) review, researchers have been prone to emphasize the importance of assessing components, sometimes going as far as to claim or imply that components must be assessed in order to address accuracy questions. Even the most avid component proponents, however, agree that obtaining the type of data necessary to do their recommended componential analysis is often extremely difficult. Although many types of research can be justifiably characterized as “difficult,” the componential approaches raised the difficulty bar to a new level, which helps explain why Cronbach’s review helped discourage researchers from addressing accuracy questions at all.

So, “must” all accuracy research assess components? Well, the answer depends on what this question means. If it means, “Must all accuracy researchers understand existing componential approaches in order to have better insights into the meaning of the results obtained from studies that do not explicitly assess components?” then my answer is “yes.” It certainly behooves all of us interested in accuracy research to have more, rather than fewer, insights into the potential sources of social perception and the processes leading to accurate or inaccurate judgments and, especially, of the limitations and potential artifacts that influence whatever index of accuracy we do use.

If, however, the question means, “Must all accuracy researchers perform componential analyses because otherwise their research will be completely meaningless or uninterpretable?” then my answer is an emphatic “no!” Here’s why.

Process versus accuracy, one more time. First, componential approaches provide one class of *explanations* for how a person arrived at an accurate or inaccurate judgment. Indeed, Cronbach (1955) titled his article “Processes Affecting Scores on ‘Understanding of Others’ and ‘Assumed Similarity’” (emphasis mine). Why? Because components provide information about the *processes* of judgment. And they do a good job of it.

How do I arrive at my prediction that Mike will hit a grand slam? Do I always predict that he will hit homeruns, or am I a particularly astute judge of Mike's hitting? How do I arrive at my judgment that Bertha is extroverted? Does everyone say she is extroverted? Or perhaps I always say everyone is extroverted. And why do I think African Americans are less likely to complete high school than they really are? Do I underestimate every group's likelihood of completing high school? Does everyone, including African Americans, underestimate African Americans' likelihood of completing high school? Or am I ethnocentric, underestimating *only* African Americans' success, and not my own group's success, at completing high school?

Answers to these types of questions address the processes by which people arrive at accurate or inaccurate social judgments. So, components give valuable insights into process.

But process is irrelevant with respect to establishing the degree of (in)accuracy of some perception. If I say, "Mike is going to hit a homerun" and he does, this particular prediction is right. End of discussion regarding my *degree* of accuracy.

With respect to understanding *how* I arrived at that prediction, it would be valuable to estimate my elevation, stereotype accuracy, differential elevation, and differential accuracy (if you like Cronbach), or, if you prefer SRM, my elevation, my perceiver effect, Mike's target effect, and our interaction effect. But if you want to determine *whether* my prediction is accurate, the only thing we need to do is figure out whether he hit the ball over the outfield fence, in fair territory.

If Vlad believes that Armenians are public parasites burdening the financial community with their constant need for charity, and Armenians actually make fewer demands on public charity than other groups (LaPiere, 1936), then Vlad overestimates the financial burden created by Armenians. Again, period, the end—at least "period, the end" with respect to *establishing* the inaccuracy of Vlad's belief.

Interpreting that inaccuracy is another matter. As Ryan, Park, and Judd (1996) have pointed out, in the absence of their full accuracy design (a research design that permits assessments of all the components in their model), we cannot conclude, as did LaPiere (1936), that this means that Vlad is necessarily a raging anti-Armenian bigot. Perhaps Vlad overestimates *every group's*, including his own group's, need for charity. In that case, Vlad is not ethnocentric at all. People with nasty beliefs about all groups, including their own, may be, well, nasty people. But if their beliefs are equally nasty about all groups, including their own, they are not ethnocentric. So, Judd and Park's (1993) full accuracy design would be extremely useful for providing some insight into *why* Vlad overestimates Armenians' request for charity. But it is completely irrelevant with respect to establishing *whether* Vlad overestimates their requests for charity. That question can only be answered by comparing Vlad's estimate of their need for charity to some criteria.

So, *establishing* (in)accuracy is a very different animal than *explaining* (in)accuracy. Establishing (in)accuracy merely involves comparing the perception (judgment, prediction, expectation, etc.) to the criteria. The more closely the perception corresponds with the criteria, the more accurate the perception.

No reification of components! I think it is extremely tempting to reify componential approaches to accuracy. First, they are presented with a sort of heavy-handed statistical rigor that gives them a veneer of being more scientific than the rest of us statistically backward folks could ever aspire to. Second, they really do capture important, fundamental aspects of

social perceptual judgment processes. Third, they successfully identify sources of bias or noise in judgments that few of us usually mean by accuracy. Thus, it is very tempting to view components as concrete fixtures on the social perceptual landscape. If they are there, then we should have to assess them, shouldn't we?

Such absolutist positions regarding components ("Cronbach's [or SRM] components must always be assessed" or "Accuracy can only be viewed componentially," etc.) are, in my view, unjustified for several reasons. First, there is no one right way to divide up components of social perception. This should be clear from my brief review of Cronbach's, Kenny's, and Judd and Park's componential approaches. They have important similarities, but, obviously, there are also differences between all three. Such differences are made salient when the three approaches appear side-by-side, as they do in Table 12-1. If there was any single "right" set of components that "must" be examined, and if components were actually hard and fast fixtures in the social perception landscape, there could not possibly be three different breakdowns of components, unless one breakdown is "right" and the other two are "wrong" or unless each was woefully incomplete.

TABLE 12-1

Componential Approaches to Social Judgment

Cronbach's (1955) components:

Judgment of a person's trait =

Constant (elevation) + Target main effect (differential elevation) +

Trait main effect (stereotype accuracy) + Target \times Trait interaction (differential accuracy)

Social Relations Model Components (e.g., Kenny, 1994):

Judgment of a person's trait =

Constant (elevation) + Perceiver main effect (perceiver effect) + Target main effect (target effect)

+ Perceiver \times Target interaction (relationship effect)

Judd & Park's (1993) Components for Research on Stereotype Accuracy:

Judgment of a group's traits =

Constant + Perceiver group effect (pge) + Target group effect (tge) +

Attribute (stereotypic vs. counterstereotypic) effect (ae) + (pge \times tge) + (pge \times ae) +

(tge \times ae) + (pge \times tge \times ae)

Hypothetical componential approach combining Cronbach's, Kenny's, and Judd & Park's approaches:

Judgment of a person's trait =

Constant + Perceiver group effect + Target group effect + Attribute (stereotypic vs.

counterstereotypic) main effect + Target main effect + Perceiver main effect + Individual

trait effect + 15 Two-way interactions +

All 20 three-way interactions + All 11 four-way interactions +

All 4 five-way interactions + The six-way interaction^a

^a This model assumes all main effect terms are independent, which may not be the case. For example, the target group main effect may not be independent of the target main effect and the attribute main effect may not be independent of the individual trait effect. In such a situation, there might be fewer interactions than displayed in this model. There would, however, still be literally dozens of such interactions. No researcher has ever advocated this model, including me.

If all were partially right but incomplete in that they failed to address components identified by other researchers, then a full componential model would need to assess *all* the components identified by all models. Such a model is presented at the bottom of Table 12-1. If components are “real” and “must” be assessed, then the only complete way to do it would be to assess the more than 50 components identified in this model. Such a model has never been recommended even by advocates of componential approaches and is not being recommended here. Indeed, it is so extreme as to border on absurd. But such an absurd model might be required if all components “must” be assessed.

The situation, however, is far more complex than even this hypothetical combined componential model suggests. There is a potentially infinite number of ways in which social perception could be broken down into components (see also Kruglanski, 1989). Attributes could be further broken down into a variety of types or subclasses (e.g., positive vs. negative; explanations vs. descriptions vs. predictions; behaviors vs. traits; and so on). Similarly, both perceiver and target groups could be broken down, not only by in-group and out-group, but by any of the infinite ways of identifying groups (culture, demographic characteristics, memberships in organizations, professional expertise, etc.).

This is not meant to suggest, however, that existing componential approaches are purely subjective and arbitrary and, therefore, can be ignored. But the choice of components will depend entirely on the types of processes one would be interested in studying and the types of response biases one would like to assess or eliminate. Different componential breakdowns serve different purposes and provide insights into different aspects and processes of social perception. Thus, understanding existing componential approaches would seem crucial to anyone studying accuracy to gain insights into how best to interpret their own or anyone else’s data addressing the degree of (in)accuracy in social perception.

Componential models may be most important when the criteria are self-reports and self-perceptions. Although Cronbach’s componential approach never generated much empirical research, Kenny’s and Judd and Park’s have, and much of that research has used target individuals’ or target groups’ self-perceptions as criteria (e.g., Kenny, 1994; Ryan, 1996). I am reluctant to use absolutes (e.g., “all” research on accuracy must be based on components), but I come awfully close when the criteria are self-perceptions, especially self-perceptions regarding traits, attitudes, or dispositions, rather than behaviors or other objective characteristics.

Self-perceptions of traits, for example, typically have no objective referent. How extroverted is someone who rates him- or herself “5” on a 1 to 7 scale with endpoints labeled “not at all” and “very”? It is hard to say because each choice is subjective, in that each rater imputes his or her own meaning to each scale point (e.g., Biernat, 1995). Such differences in subjective meanings cloud the assessment of accuracy. Componential approaches, however, are particularly good at identifying differences in subjective meaning and removing them from estimates of accuracy (this is often captured by the various elevation components). Thus, it is probably a good idea to use a componential approach, if possible, almost any time one uses self-perceptions as criteria.

Noncomponential Approaches to Assessing Accuracy

It has just been argued that componential processes are not necessary for the assessment of accuracy. This argument, however, does not rest solely on a critical evaluation of the claim

that “all accuracy research must perform a complex componential analysis.” Instead, much of the best evidence for the idea that componential analyses are not strictly necessary comes from the many noncomponential approaches to the study of accuracy that have made important and enduring contributions to understanding social perception. The next section, therefore, briefly reviews three of the most influential noncomponential approaches.

The term “noncomponential” here is potentially misleading, because it unfortunately implies that one can completely ignore the components issue. Even “noncomponential” approaches can themselves be considered to assess *subsets* of components in the various componential models, as shall be made explicit in the discussion that follows. Furthermore, componential and noncomponential models are not necessarily mutually exclusive or antagonistic; indeed, one can even take a componential approach to applying ideas from each of the two noncomponential models described below (Kenny, West, Malloy, & Albright, 2006). Nonetheless, I use the term “noncomponential” to refer to approaches that assess accuracy without an *explicit and intentional* assessment of components. When, from a componential standpoint, such approaches only assess a subset of components in one or more of the componential models, this is pointed out explicitly below. What may be *lost* by not performing a full componential analysis is also explicitly discussed.

Correlational Approaches

Most noncomponential approaches to assessing accuracy, or processes underlying accurate and inaccurate social perceptions, use Pearson's correlations to assess the extent to which judgments correspond to criteria. In general, when judgments concern a single attribute, correlations between judgments and criteria capture Cronbach's (1955, p. 191) differential accuracy correlation, which he described as: “. . . sensitivity to individual differences. . . . These are the only processes included in present measures of social perception which depend on J's [perceivers'] sensitivity to the particular O [target].”⁶

The simplest and most typical form of correlation in accuracy research is that between a set of perceivers' judgments or predictions regarding a single trait or attribute of a set of targets. For example, teachers predict students' achievement, interviewers may evaluate a set of interviewees, or perceivers may estimate the percentage of people belonging to various demographic groups that complete college. Such correlations automatically remove the elevation and stereotype accuracy components from correspondence between perceivers' judgments and the criterion. (A brief aside for the statistically inclined: This is because correlations reduce all data to deviations from the mean.) Thus, a simple correlation (between judgment and criterion) goes a long way toward eliminating many of the biases, artifacts, and problems in assessing accuracy first identified by Cronbach.

Of course, the correlation coefficient is not perfect. First, it removes or avoids, but does not directly assess, elevation and stereotype accuracy. Because correlations remove average differences between judgments and criterion, they cannot assess any consistent tendency to over- or underestimate targets (elevation, *ala* Cronbach). If it was important to assess those components in order to address some research question, one could not use the correlation to do so.

Second, correlations equate the variability of judgments and criterion. Therefore, they cannot assess whether perceivers consistently over- or underestimate target variability.

Because mean and variability differences between judgments and criteria probably often reflect response bias and/or scaling discrepancies between perceiver and criterion, these limitations to correlations do not greatly undermine their utility in assessing accuracy. I use the term “scaling discrepancies” here to refer to the idea that people may use scale points in a manner differently than used in the criterion. This would obviously be true if, for example, judgment and criterion are assessed in different metrics (e.g., subjective rating scale and percentages, respectively).

People, however, still might use the numbers in some scale differently than is used for the criterion, even if they are supposedly on the same scale. For example, let’s say Alfred estimates three people’s IQ scores as 40, 50, and 60, when they are really 115, 120, and 125. Although it is possible that Alfred believes all three of these fairly intelligent people are classifiably retarded, it is more likely that Alfred does not fully understand how IQ scores are scaled. He dramatically underestimates people’s IQ in absolute terms, but his estimates are also *overly* sensitive to actual variations in IQ (Alfred’s judgments go up 10 IQ points for every 5-point increase in actual IQ). But given his *subjective* IQ scale, the correlation between his judgments and actual IQ would be perfect (1.0), because (1) mean differences in judgment criteria are irrelevant to computation of the correlation, (2) the correlation coefficient is computed after statistically equating the variability in judgment and criterion, and (3) his judgments move in (differently scaled) lockstep with targets’ actual IQ.

Thus, the correlation coefficient would yield a conclusion that Alfred is an excellent judge of people’s intelligence. Is the conclusion justified? As long as you keep in mind that what this really means is that “Alfred is very good at detecting differences in people’s intelligence, but does not tell us anything about either how Alfred uses the IQ scale or about whether he consistently over- or underestimates people’s intelligence,” it is perfectly justified.

Construct Validity and Accuracy

In Chapter 10, I argued that assessing accuracy was much like assessing the validity of many social science constructs. This is important here, because the correlation coefficient is so frequently used to establish the validity of some measure that it is often referred to as the “validity” or “validity coefficient” of some measure (e.g., Campbell & Stanley, 1963; Cook & Campbell, 1979; Dawes, 1979). In much the same manner, correlations can be used to establish the accuracy of social perception. Because establishing accuracy is in many ways so similar to establishing construct validity, I next briefly review some of the main ideas underlying social science approaches to construct validity.

Basic construct validity. An extended discussion of the richness and complexity of establishing validity of some measure, construct, theory, etc., is beyond the scope of this chapter (see, e.g., Campbell & Stanley, 1963; Cook & Campbell, 1979; Cronbach & Meehl, 1955, for such discussions). However, the basic ideas can be summarized by the duck test described in Chapter 11 (if it walks like a duck. . .).

How do we know that people even “have” self-esteem, intelligence, attitudes, personality, etc., if we cannot directly observe them? That is where the issues of constructs and construct validity come in. A construct is, in essence, a mini-theory regarding the existence of some phenomenon. It includes, at minimum, a definition of the phenomenon and some clear

hypotheses regarding how that phenomenon should manifest. Construct *validity* refers to ways of demonstrating that some construct really does work as hypothesized and of ruling out alternative explanations.

This has all been very abstract so far. What does it mean to “show that a construct works as hypothesized”? Well, it could mean lots of things in different situations. Consider a construct regarding the existence of a psychological attribute, trait, or characteristic, such as intelligence, self-esteem, depression, ideology, etc. Establishing the validity of some method of measuring that psychological construct (e.g., self-report questionnaire, reaction time task, etc.) often might mean something like assessing the relationship of that measurement method with (1) other people’s agreement about a target’s possession of the attribute being measured, (2) target behavior that should reflect that attribute, and (3) target responses on some sort of standardized test. If all of these observed measures converge reasonably well, then we have probably done a pretty good job of establishing the validity (likely existence, probabilistic reality, etc.) of the underlying attribute and our means of measuring it.

Consider intelligence. Maria’s brilliance might lead at least some people to believe in her brilliance, it might lead her to engage in some highly intellectual activities, and it might lead her to receive high scores on standardized intelligence and achievement tests. Indeed, highly sophisticated statistical methods have existed for some time now for estimating the extent to which underlying, unmeasured attributes predict observed variables that are supposed to reflect those underlying attributes (e.g., Bollen, 1989; Joreskog & Sorbom, 1983).

If, for example, everybody says Maria is really smart, and Maria spends her time reading Einstein’s original works, and she scores at the high end of a test that is supposed to measure intelligence, then we have probably fairly well-established both the utility of the intelligence construct and the fact that she is pretty smart. (Obviously, we would need to do this for more than one person and it would need to work just as well on the low and middle portions of the intelligence spectrum as on the high end, but I hope you get the idea).

Consider extroversion. Let’s say everyone describes Ian as a wild and crazy guy, we find that he spends much of his spare time attending or holding parties, and he scores highly on a personality test assessing extroversion. It sure is beginning to look like extroversion may be a (probabilistically) real attribute, and Ian scores highly on it (as with intelligence, we would need to do this for more than one person and show that this works just as well for people low or average in extroversion).

Thus, intelligence and extroversion are both *constructs*, but, in the examples above, constructs that have been reasonably well-validated. As such, we now have license to treat them as real—with one big caveat. It is always possible that someday someone will come along with evidence that questions, challenges, or even successfully undermines the validity of either the construct or some previously well-established measure of the construct. Until that time, however, it is reasonable to act as if the construct is about as real as apple pie, and most researchers will treat it that way (if it looks like a duck. . .).⁷

Accuracy. Construct validity is very nice, but where is the accuracy? Just because scientists can, within the context of probabilistic realism, identify that Maria is smart or that Ian is extroverted through a variety of rigorous scientific methods does not necessarily mean that regular everyday walking around people are necessarily very accurate.

This is true. Establishing accuracy involves establishing correspondence between perception and reality, not just establishing that some attribute (intelligence, extroversion, tennis

ability, etc.) can be successfully measured. Even statistical beginners, however, should know how to establish “correspondence” between two variables—just correlate them! If my beliefs that Maria is smarter than Ian but Ian is more outgoing than Maria are reasonably accurate, then those beliefs should correlate well (not necessarily perfectly!) with their behavior reflecting intelligence and extroversion, with others’ beliefs about their intelligence and extroversion, and with their scores on tests assessing intelligence and extroversion.

Thus, establishing accuracy in social perception for correlational approaches is highly similar to establishing construct validity. But this means that establishing accuracy is not much more difficult than establishing construct validity! Establishing construct validity is not easy—it is sufficiently complex that whole books have been written about it. Nonetheless, it suffers from none of the controversy, hand wringing, or intellectually imperious dismissiveness regarding supposed conceptual, political, or criterion “problems” one often finds in the literature criticizing accuracy (see Chapters 10 and 11 for reviews of such controversies).

Establishing accuracy, however, is a bit trickier than establishing construct validity for several reasons. First, self-fulfilling prophecy and bias have to be ruled out as explanations for the correspondence between belief and criteria, although this is not easy, nor is it inordinately difficult, either (and Chapter 10 discussed some of the many ways of doing so).

Second, establishing accuracy is a bit trickier than establishing construct validity because social perception, judgment, and expectations are themselves constructs! You cannot feel someone’s judgment or taste their expectations. Thus, all the issues involved in establishing construct validity kick in, not just when measuring targets’ attributes, but when measuring perceivers’ expectations and beliefs about others.

Accuracy, therefore, in this imperfect world where little can be known with certainty and observed measures are only probabilistically related to underlying attributes, *is not* usually best reflected by correlations between *observable* measures (e.g., a measure of perceiver expectations and a measure of target extrovertedness). Accuracy will often best be reflected by correlations between the underlying constructs representing the social perception and the criteria (for the statistically inclined, this can often be accomplished either by disattenuating correlations for unreliability or by using LISREL-type models—see, e.g., Bollen, 1989; Carmines & Zeller, 1979).

Thus, assessing relationships between underlying constructs for expectations and criteria will usually yield the best estimate of accuracy. This should not, however, be misinterpreted to mean that all accuracy research must necessarily assess relationships between underlying constructs rather than observed measures. Although doing so will usually provide the best assessment of accuracy, sometimes it may just not be possible. In such cases, more information regarding accuracy and inaccuracy would be provided by assessing correlations between observed measures of underlying attributes, judgments, or expectations, rather than by assessing nothing at all. Such correlations will tend to underestimate accuracy to the extent that the observed measures only imperfectly reflect the underlying attributes or expectations. This, of course, does not constitute any sort of immovable obstacle to or fatal flaw in accuracy research. It simply means that people may be more accurate than indicated by research that only assesses correlations between observed measures of expectations and observed measures of criteria.

This brief delving into the statistical and methodological arcania of construct validity and unmeasured variables was necessary to lay the foundation for understanding three of the

main *noncomponential* approaches for assessing both degree of (in)accuracy and processes underlying (in)accuracy in social perception. All three are fundamentally based on the correlation between perception and criteria.

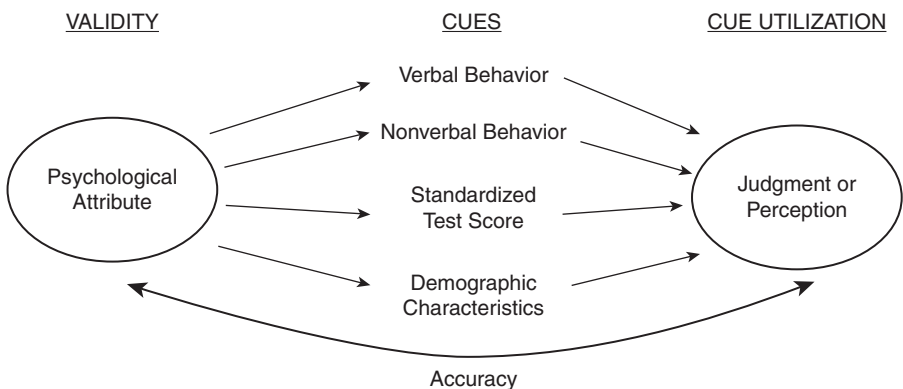
Brunswik's Lens Model

Brunswik (1952) suggested that accurate perception of reality (both object and social) involves the use of cues probabilistically related to objective reality. He metaphorically called his approach the Lens Model to capture the idea that objective reality is never observed directly. Instead, cues related to objective reality must be observed and interpreted as relevant to some judgment—that is, objective cues are seen through the “lens” of subjective perception and judgment.

This does not necessarily mean that perception is a purely subjective phenomenon unrelated to objective reality. Indeed, one of the main purposes of the Lens Model is to provide a mechanism not only for assessing people's degree of accuracy but also for understanding sources of both accuracy and inaccuracy in their judgments.

Figure 12–1 presents a simplified but general version of the Lens Model. It captures several main ideas. First, the circled “Psychological Attribute” represents some sort of psychological construct that cannot be directly observed (self-esteem, extroversion, intelligence, etc.). The Cues, shown in the middle of the figure, are directly observable or measurable phenomena. The arrows pointing from the Psychological Attribute to the Cues are labeled “Validity,” because they represent the extent to which the underlying attribute manifests itself in the observable Cues.

The rightmost circle represents perceivers' judgments (or perceptions). The arrows going from the Cues to Judgments represent the extent to which the observable cues influence perceivers' judgments (labeled “Cue Utilization”).



Single-headed arrows are causes the double-headed arrow is a correlation. This figure presents an example of how the Lens Model might be used. It can be used for assessment of physical attributes (e.g., size, distance); as well as of psychological attributes. The four cues here are concrete examples of potential manifestations of some attribute: the Lens Model is not restricted to these cues. In Lens Model research, what is called “accuracy” here is usually called achievement.

FIGURE 12–1 Brunswik's Lens Model

Thus, the Lens Model characterizes social perception as a two-step process: (1) observable manifestation of psychological attributes and (2) perceiver use of observable cues to arrive at judgments. Accuracy, therefore, is captured by the correlation between the psychological attribute and the judgments (the long, double-headed arrow in Figure 12-1). Correlations assess how well the judgments correspond to the attributes—i.e., accuracy.

The Lens Model is a noncomponential, correlational model for assessing both degree of accuracy and processes of social perception. Identifying cue validity and cue utilization focuses on a very different set of the processes than is typically the focus in componential models. As such, it provides different (not better or worse) types of insights into processes of social perception than do componential models.

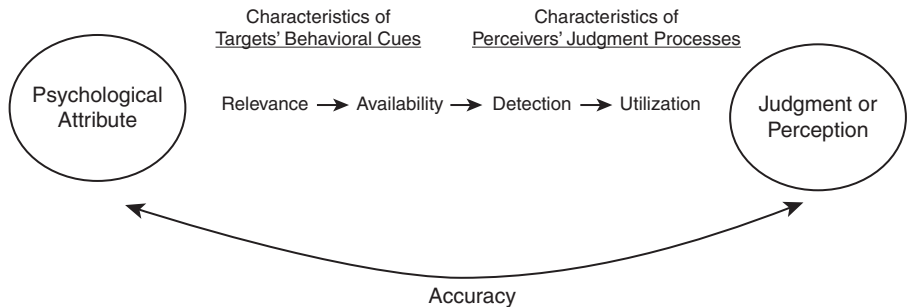
The Realistic Accuracy Model

Funder's (1995) Realistic Accuracy Model (RAM) draws on essentially the same set of fundamental assumptions I described under "probabilistic realism" in Chapter 11 (indeed, Funder's perspective inspired much of that section) to create a model that could be viewed as an extension and elaboration of Brunswik's Lens Model. Some of the main ideas of RAM are depicted in Figure 12-2.

As with the Lens Model, overall accuracy is typically assessed by the correlation of the underlying attribute with the perceiver's judgment (represented by the large, curved, double arrow on the bottom of Figure 12-2). Four steps needed for perceivers to arrive at an accurate judgment are displayed in between the underlying attribute and the judgment.

First, the underlying attribute needs to create some sort of observable evidence relevant to that attribute (the cues, in the Lens Model). Dishonesty, for example, is not likely to be displayed in a large lecture hall (except perhaps during test time). Interest in the class is more likely than honesty to be displayed, for example, through high attendance levels, keeping up with class assignments, and/or class participation.

Second, the evidence needs to be at least hypothetically available to the perceiver. Whether or not a student has completed all assigned readings may rarely be directly observable to the class's teacher. Attendance and participation, however, would be considerably more available.



Single-headed arrows here represent steps in a sequence, not causal effects. The double-headed arrow represents the correlation between targets' attributes and perceivers' judgments.

FIGURE 12-2 Funder's Realistic Accuracy Model

Third, the perceiver has to detect that evidence. In a large lecture hall, which students participate in is pretty obvious. However, detecting precisely which students do and do not attend regularly, out of a swarming mass of hundreds of students in the lecture hall, is considerably more difficult (unless attendance is somehow recorded).

Fourth, the perceiver has to actually use the detected evidence/cues for arriving at a judgment. If lecture hall teachers can't remember which students regularly participate, it would be pretty difficult to use participation as a basis for judgment regarding interest in the class.

Although this is the heart of Funder's (1995) RAM, applying the model might be considerably more complex. People have lots of underlying characteristics. Funder (1995) focused primarily on personality traits, but RAM seems applicable to all sorts of unobservable personal characteristics, including, for example, emotions, attitudes, motivation, etc. Furthermore, one attribute may create many cues (as is most obviously captured in the Lens Model), and some cues may reflect multiple attributes. Thus, one type of complexity involves the sheer number of possible interrelationships among attributes, cues, and judgments.

Like several of the componential models, RAM also considers how the perceiver, target, attribute, and evidence relate to accuracy (Funder [1995] referred to these as Judge, Target, Trait, and Information, respectively, but I am sticking with my terminology here). But this is not to identify components. Instead, the purpose is to analytically identify how specific combinations of perceiver, target, attribute, and evidence might combine to influence accuracy.

For example, some perceivers may be particularly good (poor) at evaluating certain types of traits (e.g., clinical psychologists might be better than most of the rest of us at evaluating others' mental health). Some perceivers might be particularly good (poor) at judging particular targets (e.g., close friends might be better judges of each other than are strangers). Some perceivers might be especially good (poor) at using or obtaining certain types of evidence (e.g., some people may be better than others at picking up on targets' emotion-revealing nonverbal cues). Funder's (1995) article goes into considerable length regarding how the various combinations of perceiver, target, attributes, and evidence may combine to influence accuracy.

Like the Lens Model, RAM assumes that relationships between underlying attributes, cues, and judgments are probabilistic. Like the Lens Model, overall accuracy is typically assessed via correlations, although discrepancy scores (between judgment and criterion) can be used, too (see Funder, 1987, 1995). RAM is particularly good at explaining why accuracy in person perception may often not be all that high. For the judgment to closely correspond to the criterion, that criterion needs to clearly manifest itself in ways that could be, and in fact are, detected by the perceiver, and then the perceiver needs to use that detected information (as well as not use information that is not relevant to the judgment). A breakdown at any step will dramatically undermine accuracy. Furthermore, by focusing on combinations of perceiver, target, attributes, and evidence, RAM is also particularly good at highlighting processes that may enhance or undermine accuracy.

Dawes' (1979) Improper Linear Models

Dawes (1979) made a very interesting discovery. In reviewing his own and others' research on decision making, he discovered that (1) people tend to be very good at identifying the

evidence or cues that are relevant to making some prediction, but (2) they are not very good at combining or integrating those cues. Thus, their overall predictive accuracy tends to be quite low. Note, however, that this is not because people are completely out to lunch (biased, error-prone, etc.). They are good at one part of the prediction task (identifying criteria for making a prediction) but poor at another part (putting those criteria together).

Consider admissions to graduate school in psychology. The criteria typically used for making admissions decisions seem appropriate: GRE scores (general intellectual ability), GPA (achievement at academic tasks over an extended period), and letters of recommendation (what experts in the field who are highly familiar with the applicant have to say about him or her). Nonetheless, Dawes (1979) found that the correlation of graduate admissions committee evaluations with later success in graduate school is typically quite low (.19).

If people are completely out to lunch, then they would not even use appropriate criteria—that is, the criteria they do use would not predict success in graduate school. However, if they are good at identifying the appropriate criteria but use them poorly, then the raw criteria themselves should do a much better job at predicting graduate success. This was indeed the case—the overall (multiple) correlation of the criteria themselves with graduate success was about .4.

What to do? It is unreasonable to expect admissions committees to compute complex statistical formulas in their heads or to create a formal statistical score for each applicant. Dawes provided an elegantly simple and even amusing answer. Identify the criteria, weight them all equally, and add. For example, GRE, GPA, and letters of recommendation might each be transformed onto a 1 to 10 scale.⁸ Priscilla, with good GREs, a high GPA, and excellent letters of recommendation, might receive weights of 7, 9, and 9, respectively, for a total score of 25. George, with high GREs, a good GPA, and good letters, might receive scores of 9, 7, and 7, for a total of 23. Priscilla would be ranked more highly than George.

This is different from a formal statistical model primarily because the weights for each predictor have been chosen in a less than optimal manner (many statistical prediction techniques, such as regression, identify how to weight the criteria in such a manner as to maximize their overall predictive validity). But here is the second amazingly elegant aspect of Dawes' analysis: Equal-weight, easy-to-compute, improper linear models predict outcomes nearly as well as do formal statistical models! In the graduate admissions example, Dawes' improper linear model correlated .38 with future success in graduate school. Dawes (1979) went on to show that a simple, improper linear model performed similarly well in predicting all sorts of outcomes, including choice of bullet type for a police department and a bank's predictions regarding companies likely to go bankrupt.⁹

Dawes' improper linear model is fundamentally different than the Lens Model and RAM. The Lens Model and RAM were specifically designed to assess degree of accuracy and processes underlying social perception. That is, they were meant to *describe* aspects of the social perception process. In contrast, Dawes' model is primarily prescriptive (it suggests how people should go about making decisions and arriving at predictions).

Nonetheless, I have included it here for two reasons. First, Dawes' (1979) conclusion that people are good at selecting criteria but not good at using them is descriptive. In RAM terms, it suggests that people often are good at detecting available and relevant cues but that they often do not utilize them well (in Lens Model terms, their cue utilization would be poor). Second, although Dawes did document that people were, on their own, not very good

at arriving at accurate predictions, he also showed that the accuracy of their predictions could easily be improved. Identify the criteria, weight them equally, and then add!

Noncomponential Models: Final Comments

Correlational approaches to accuracy, including but not restricted to the Lens Model and RAM, do not oppose or contradict componential approaches. Indeed, it is quite possible to perform a Lens Model or RAM analysis via components, if one felt that would be useful or important (Kenny et al., 2006). Furthermore, simple correlations between criterion and judgment go a long way toward eliminating many of the often irrelevant artifacts and biases identified by Cronbach (1955) and Kenny (1994). Nonetheless, my point for presenting them here has not been to argue that they refute componential approaches. My point, instead, has only been to demonstrate that some very sophisticated and successful noncomponential models and approaches to accuracy have been developed. One will find no mention of components in Brunswik (1952) or Funder (1995), or in many other influential articles and books on accuracy (e.g., Ickes, 1997; Jussim, 1991; McCauley, Stitt, & Segal, 1980; Swim, 1994). Components are interesting and important, but claims that one must always assess components when studying accuracy are not justified.

Accuracy: Conclusions

I have just spent three chapters discussing accuracy but have described very little empirical research assessing people's accuracy. How accurate are interpersonal expectations? Not answered (yet). Do teachers' expectations predict student achievement primarily because of self-fulfilling prophecy or accuracy? Not answered yet. How accurate are people's beliefs about demographic groups? Despite three chapters on accuracy, I have still not reviewed research assessing the actual (in)accuracy of social stereotypes.

Why not? For several reasons. First, assessing accuracy is a genuinely complex undertaking and is also theoretically and politically controversial. Therefore, I felt it was necessary to explore some of those complexities and controversies before describing the relevant research findings. Second, in my opinion, those complexities, although real, have often been characterized as "problems" or "difficulties," and once characterized as such, have led many social scientists to despair at the prospect that accuracy can even be assessed (or to denigrate the value of attempting to do so).

Chapter 10 was necessary to review and critically evaluate many of the historical reasons for the demise of accuracy research in social psychology. In many cases, it contested the viability of many of the common criticisms of accuracy research, concluding that such criticisms were themselves often more flawed than accuracy research itself. In other cases, Chapter 10 concluded that even the most valid of those criticisms only warranted care and caution in interpreting accuracy research, rather than a wholesale dismissal of the entire accuracy endeavor.

Chapter 11 explored the crucial issue of identifying criteria for establishing accuracy. This chapter was necessary to (1) demonstrate that, although many social cognitive process-oriented researchers have suggested that identifying criteria for establishing accuracy is so difficult as to cast a significant cloud over the viability of accuracy research (e.g., Jones, 1985;

Fiske, 1998; Stangor, 1995), the logic of establishing accuracy of social perception overlaps almost completely with the (minimally controversial) logic of establishing construct validity in the social sciences; and (2) identify myriad useful potential criteria for assessing the accuracy of social judgments.

The purposes of the current chapter have been to (1) present a simplified review of the ideas underlying the three main componential approaches to accuracy; (2) argue that, although componential approaches are valuable and important, not all research on accuracy must necessarily assess components; and (3) review some of the major noncomponential approaches to assessing accuracy and social perception processes.

Thus, Chapters 10, 11, and 12 were necessary to convey the scientific foundation on which accuracy research rests. There is, however, a third reason I have not yet reviewed much accuracy research. The best way to draw general conclusions about the relative roles of accuracy, self-fulfilling prophecy, and bias in interpersonal expectations is to perform research that assesses at least two of these three expectancy phenomena and, preferably, all three. Such research is in a much better position to place evidence regarding the power and pervasiveness of each phenomenon in context. Research that only assesses one phenomenon at a time, such as most of the research on self-fulfilling prophecies reviewed in Chapter 4 and the research on bias reviewed in Chapter 5, cannot possibly provide direct information about the *relative* roles of accuracy, self-fulfilling prophecy, and bias in interpersonal expectations.

Indeed, research focusing on only one phenomenon at a time is potentially extremely misleading. Consider 10 hypothetical studies all finding statistically significant evidence of perceptual bias. When considered together, but in isolation from research on accuracy, they might seem to support a conclusion emphasizing the almost universal presence of bias. After all, all 10 studies found evidence of bias. It sure looks like people are biased almost all the time. And from this conclusion, only a tiny step further gets you to the conclusion that bias pervades social perception. (Although 10 studies demonstrating accuracy here would be just as potentially problematic, the 10 studies demonstrating bias situation more closely mirrors the reality of social psychological research and conclusions regarding bias—see Chapters 5, 9, and 10).

But such a conclusion could not be justified by research that only examined bias. Research that does not test for accuracy cannot possibly provide any evidence of low or high accuracy. Thus, 10 studies demonstrating bias could not rule out the possibility that people tend to be far more accurate than biased in their social judgments. Combine the tendency to overinterpret significant evidence of bias with insufficient consideration of the effect size associated with bias (compare, e.g., the conclusions regarding the power and pervasiveness of both bias and self-fulfilling prophecy highlighted in Chapters 4 and 5 with the evidence regarding self-fulfilling prophecy and bias summarized in Table 6-1 in Chapter 6) with the almost complete exile of accuracy research from social psychology from 1955 to 1985, and it becomes easy to understand much of the source of social psychological emphasis on bias and self-fulfilling prophecy.

For the same reason, therefore, I will not present a chapter devoted to research that exclusively focuses on accuracy. I have no objection in principle to research focusing exclusively on accuracy (or on bias or on self-fulfilling prophecy). However, all three phenomena (bias, self-fulfilling prophecy, and accuracy) characterize interpersonal expectations. Furthermore, social psychology's history of emphasizing self-fulfilling prophecy and bias relative to accuracy seems to result more from researchers' emphasis on studying those phenomena than

from actual empirical results demonstrating that self-fulfilling prophecy and bias are high relative to accuracy. On top of that, lines of research that only focus on one of these phenomena at a time can even be viewed not as “talking to each other” but as “talking past each other” (e.g., social psychological studies and reviews of interpersonal expectations have almost never cited educational research demonstrating high accuracy in teacher expectations). Therefore, in this book, I bring these separate lines of research together in order to provide a fuller, more balanced, and more valid perspective regarding the extent to which expectations create versus reflect social reality.

Just as research focusing exclusively on biases or errors implicitly conveys the idea that social perception is dominated by flaws, a section focusing exclusively on accuracy might implicitly convey the idea that social perception is nothing but accurate (a view that is also not supported by the evidence). Instead, the next chapter is titled “Accuracy and the Quest for the Powerful Self-Fulfilling Prophecy” in order to present a much more even-handed vision of the extent to which expectations are accurate versus create self-fulfilling prophecies or biases. Doing so, however, requires discussing these phenomena together, rather than in isolation from one another.

Notes

1. For those of you with some (but not a lot of) familiarity with baseball, having a player on second or third base is referred to as “runners in scoring position,” because a single will usually drive in a run. This includes having a runner on second, having a runner on third, having a runner on second and third, having runners at first and second or at first and third, and having the bases loaded. For those of you completely unfamiliar with baseball, this will not help at all.

2. Cronbach's components actually involve comparisons of ratings, not rankings. However, once you start removing components, what's left is not really a rating, either. It is a sort of “rating adjusted for removal of other components.” Rankings capture most of what's important from such “adjusted ratings.” That is, the main accuracy question is something like: “Does the rank order of adjusted judgments correspond to the rank order of adjusted criteria?” Therefore, I also refer to rankings when discussing stereotype accuracy and differential accuracy.

3. If you do not know anything about baseball, with a little help, this example should still be pretty easy to follow. “Hits” are good, at least if you are the hitter. Throughout a game, each player usually has several opportunities to try to get a hit. For each player, one can compute a batting average, which is simply the proportion of times that player has succeeded in getting a hit. For example, a player who had 3 hits in 10 chances would have an average of .300 (3/10).

4. For those of you unfamiliar with baseball: The pitcher is on the team opposing the batter. The pitcher's job is to get the batter “out”—that is, not allow a hit.

5. For those of you uninitiated in baseball, here are the key elements of this situation. If George gets a hit, the runners on second and third will likely score, and his team will win the game 6–5. But if Lillian can get George “out,” the game will be over and Lillian's team will win.

6. Cronbach was referring to, in addition to differential accuracy correlation, differential elevation correlation in this quote, which refers to the correlation between the perceivers' judgment of a target averaging over all attributes and the targets' average score on the criteria representing those attributes. This, however, is irrelevant when there is only a single attribute being judged.

7. In their practice of inferring the existence of unobserved phenomena from their theoretically predicted observed effects, social scientists are in some pretty good company. Physicists have never seen a neutron, proton, electron, or quark; cosmologists have not witnessed the big bang; and Darwin never witnessed speciation. Like psychological attributes, the existence of these unobserved phenomena is inferred from their effects.

8. For those of you with an even introductory familiarity with statistics, there is an even better alternative: Standardize all predictors, and then add. This weights all predictors equally by virtue of not only putting all variables on the same scale but also equating their variances. I thank Gretchen Chapman, Rutgers' resident expert on decision making, for this suggestion.

Such procedures may work, in part, because they reduce reliance on salient, easy-to-use criteria. In college and graduate admissions, for example, I suspect that most programs rely very heavily on standardized test scores, such as the SATs, GREs, MCATs, LSATs, etc., not because someone has made the decision that these should be the main criteria (indeed, it is probably easier to find righteous proclamations denying that admission is based primarily on these criteria than to find defenses of the appropriateness of doing so), but because using such numbers is so easy (as compared to, e.g., letters of recommendations, personal statements, and even GPA). A Dawesian simple, improper linear model that includes standardized test scores as only one criterion among many goes a long way toward eliminating any tendency to overemphasize such scores in making admissions decisions.

9. Dawes' results were actually even more amazing, at least for the statistically inclined. The equal-weight improper linear model generally (i.e., not just in the graduate admissions example) predicted outcomes *better than* the split-sample cross-validated regression weights.

5 The Quest for the Powerful Self-Fulfilling Prophecy

This page intentionally left blank

13 Teacher Expectations

ACCURACY AND THE QUEST FOR THE POWERFUL SELF-FULFILLING PROPHECY

Teachers' Pets Versus the Troublemakers

Teachers' pets—students who, through charm, social skills, and an easy conformity to the rigid (and, in my opinion, sometimes ridiculous) rules of the primary or secondary school classroom—used to drive me nuts. School, I thought, was about learning and mastering material—and if one did that, regardless of how charming (or uncharming) one was, one deserved high grades. Although the pets were usually pretty smart, they never seemed to me much smarter than anyone else. What drove me nuts was the consistent manner in which they were favored by the teacher, especially when it came to grades.

Teachers' pets were most likely the beneficiaries, to at least some extent, of self-fulfilling prophecies. Of course, most pets were good students anyway, so it is difficult to tease out how much of their performance resulted from their own skills, motivation, and competence (independent of effects of special teacher treatment) versus how much they benefitted from special treatment. Nonetheless, all the warmth, positive feedback, and getting the benefit of the doubt in marginal cases probably increased not only grades but also their interest in school and motivation to succeed. It was, however, especially striking to me that the SAT scores of the teachers' pets I went to high school with (when I knew them) did not seem to be any higher than those of most of my other, less favored, classmates. Thus, if there were self-fulfilling prophecies, they did not seem to extend much beyond any particular class.

In my twelfth grade advanced biology class, there were also three “anti”-pets. All three students despised the teacher and were routinely rude, disrespectful, and disruptive. All three

also failed this class. Now, there are a few possible accounts for this. We can rule out all three being unintelligent by traditional standards, because all three scored over 1,200 on their SATs, graduated in the top 15% of their class, and went on to receive not only college degrees but also graduate degrees (two PhDs, one master's). We can also rule out that all three were generally bad at science; all three had received nothing but As and Bs in prior science classes. Furthermore, one became an economist, one a geologist, and one a psychologist. One possibility is that, even if their objective performance was not altered, this teacher's resentment of these disruptive students may have biased his evaluation of their performance.

Did the teacher's resentment also translate into negative expectations that created a self-fulfilling prophecy? It is hard to know. Certainly, one explanation for their failure is that his retaliatory hostility caused them to perform more poorly in this class. Furthermore, none pursued careers or even college work in biology. Perhaps their bad experiences in this class discouraged them from pursuing college coursework in biology and biologically oriented careers.

These stories—of the classic pets and the troublemakers—are, of course, merely stories. They are not hard, scientific research. However, they at least raise some issues relevant to understanding the role of teacher expectations in predicting and causing student achievement. First, they raise the question of the accuracy of teacher expectations. Clearly, teacher expectations are not always perfectly accurate. They probably overestimate the capabilities of some students and underestimate others. But do teachers typically hold inaccurate expectations for students? Or are wildly inaccurate expectations the exception, rather than the rule?

Second, what are the typical effects of teacher over- and underestimates of students? Do teacher inaccuracies harm students more than they help them or do teacher inaccuracies help students more than they harm them?

Third, these stories raise the possibility that biases and self-fulfilling prophecies may, at least sometimes, be quite large. Most students are neither pets nor antipets, so these stories suggest that although such powerful effects may occur, they are likely to be unusual. Nonetheless, might it be possible to identify conditions under which self-fulfilling prophecies in the classroom are systematically larger than the typical, small expectancy effect?

The Three Questions Upon Which This Chapter Is Focused

This chapter is framed around three questions that capture many of the issues raised in the pet examples and which have been addressed by the teacher expectation research that moved beyond the controversies surrounding the original Pygmalion study (Rosenthal & Jacobson, 1968a, b):

1. How powerful are expectancy effects in the classroom?
2. How accurate is the typical teacher expectation?
3. Have any conditions been identified under which truly powerful self-fulfilling prophecies do occur?

These three questions were selected for several reasons. In addition to providing a framework for reviewing some of the major themes of empirical research on teacher expectations,

each question also reflects a *modern* controversy. Although some of these issues have been discussed in prior reviews (Brophy, 1983; Brophy & Good, 1974; Jussim, Smith, Madon, & Palumbo, 1998; Rosenthal, 1974; Snow, 1995; Spitz, 1999; Wineburg, 1987), (1) none of the prior reviews has simultaneously synthesized research addressing all three questions in an attempt to reach a set of integrated conclusions (other than Jussim & Harber, in 2005, which was largely based on earlier drafts of Chapters 3, 13, and 14 from this book), and (2) the answers to those questions remain controversial today. Specifically, it is easy to find recent literature suggesting or explicitly espousing diametrically opposed conclusions regarding each question. This chapter, therefore, not only highlights how the accumulated evidence bears on these controversies but also shows how the results of that research can be integrated into a relatively small number of straightforward conclusions.

Each question, furthermore, goes to the heart of social psychological claims emphasizing the potential role of teacher expectations and self-fulfilling prophecies in social problems. This potential appears to be a central reason for some of the heat in the controversies that have enveloped this area of research from the outset. Specifically, the social problems view of teacher expectations suggests that the answers to these questions are that expectancy effects are real, widespread, and powerful; they can have profound effects on achievement; they are frequently inaccurate; and negative expectancy effects are stronger than positive ones (for reviews emphasizing the role of teacher expectations or self-fulfilling prophecies in social problems see, e.g., Claire & Fiske, 1998; Darley & Fazio, 1980; Fiske & Taylor, 1991; Gilbert, 1995; Hamilton, Sherman, & Ruvolo, 1990; Jones, 1986, 1990; Schultz & Oskamp, 2000; Taylor, 1992; Weinstein, Gregory, & Strambler, 2004). Next, therefore, I evaluate this perspective by examining how the scientific evidence bears on the questions framing this chapter.

How Powerful Are Teacher Expectation Effects?

Is this a silly, trivial, or unanswerable question? Some would argue that it is, because any absolutist statement regarding the power of teacher expectations will be disconfirmed by some evidence, and because teacher expectations produce stronger self-fulfilling prophecies under some conditions than others. Furthermore, some researchers have argued that all that one can do is identify what phenomena hold under which conditions, and then make educated guesses about their prevalence in daily life (Fiske & Neuberg, 1990).

The unjustifiability of absolutist statements, however, does not preclude the possibility of drawing broad generalizations from the large accumulated data on teacher expectations. The importance of doing so is implicitly testified to by the frequency with which researchers make broad generalizations about the power of self-fulfilling prophecies (see the quotes in Chapter 5). Why would such statements appear so often if the writers did not consider them important? And, indeed, it would be the height of double standards to suggest that it is acceptable to make broad generalizations (if the evidence warrants it) suggesting that the evidence testifies to the power and pervasiveness of error, bias, and self-fulfilling prophecy, but that it is not acceptable to make equally broad generalizations (if the evidence warrants it) suggesting that evidence testifies to the weakness of error, bias, and self-fulfilling prophecy and the power of accuracy. The evidence might more strongly justify one or another

conclusion, but generalizations, per se, cannot be acceptable when they support a cherished view but unacceptable when they do not.

Although broad generalizations permit many exceptions, they are extremely valuable for at least two reasons:

1. They provide a summary of the major patterns of findings in some domain of research. For example, if taking aspirin reduces one's risk of heart disease by 20%, this would seem worth knowing, even if some people regularly taking aspirin still develop heart disease. Similarly, if teacher expectations are more often accurate than inaccurate and typically have weak effects on student achievement, this would seem worth knowing even if there are some relatively uncommon situations in which teacher expectations sometimes have a powerful influence.
2. Understanding of the broad patterns of findings in some area should influence and constrain narrative nonempirical discussions of that phenomenon. For example, if it was discovered that taking massive amounts of vitamin C does little to increase resistance to colds and flu, then discussions extolling the virtues of vitamin C to reduce susceptibility to colds and flu would not appear justified. Similarly, if it was discovered that teacher expectations typically predict student achievement primarily because they are accurate and to only modest degrees because they are self-fulfilling or biased, conclusions extolling the power of teacher expectations to create self-fulfilling prophecies and biases would not appear justified. In this spirit, therefore, the next section discusses how the accumulated evidence regarding teacher expectations bears on the strikingly different generalizations frequently found in the educational and social psychological literatures.

There are two types of expectancy effects: self-fulfilling prophecies, in which teacher expectations change student achievement, and biases, in which teacher expectations alter their own judgments and perceptions of student achievement. Although far more teacher expectation research has addressed self-fulfilling prophecies than expectancy-maintaining biases, both are discussed next.

Self-Fulfilling Prophecies in the Classroom

Effect sizes in experimental studies. The typical or average self-fulfilling prophecy in the classroom uncovered by empirical research is, by any standard, quite small. The overall self-fulfilling prophecy effect (all effect sizes here are in terms of the correlation coefficient, r) in the original Pygmalion study (Rosenthal & Jacobson, 1968a, b) was .15. Raudenbush's (1984) meta-analysis of the effects of experimentally induced teacher expectations on student IQ showed the overall effect was near zero, and that even the most powerful systematic effects were only around .2 (obtained when expectations were manipulated early in the year and among students in first, second, and seventh grades). Both meta-analyses and narrative reviews (e.g., Brophy, 1983; Jussim, 1991; Rosenthal & Rubin, 1978; Raudenbush, 1984) support the conclusion that, on average, teacher expectation effect sizes are small (.1 to .2, in terms of correlation or regression coefficients).

Effect sizes in naturalistic studies. The great strength of experiments is their ability to assess causal relations. With respect to understanding the effects of expectations that teachers typically develop, however, experiments have several crucial limitations that severely restrict their ability to provide generalizable conclusions about the power of self-fulfilling prophecies in the classroom. First, whether teacher expectations are typically as inaccurate as those that are experimentally induced is unknowable merely from experimental research that intentionally induces false teacher expectations! If, however, naturally occurring teacher expectations are more accurate, they will have less potential to create self-fulfilling prophecies.

Second, if one wishes to reach broad and general conclusions on the basis of an experimental study, all that one can do is hope, argue, and speculate that one's methods and procedures come reasonably close to mimicking naturally occurring phenomena and processes. To reach scientifically justified conclusions about the "typical" power of self-fulfilling prophecies in the classroom, however, would seem to require at least sometimes actually assessing such power under typical conditions. At minimum, the type of artificial induction of false expectations common to experimental studies of teacher expectations (and self-fulfilling prophecies more generally) would need to be jettisoned because, for example, it is clear that very few teachers naturally develop expectations on the basis of intentionally falsified standardized test results provided by researchers from major universities. If one wants to reach conclusions about the potentially self-fulfilling effects of teachers' naturally developed expectations, one needs to study . . . the effects of teachers' naturally developed (not experimentally induced) expectations.

The clearest information regarding the power and pervasiveness of self-fulfilling prophecies in the classroom has been provided by studies using path analytic techniques, such as regression and LISREL, to assess the extent to which teacher expectations earlier in the school year predict changes in student achievement later in the school year.¹ Next, therefore, I discuss a handful of the naturalistic studies that have addressed self-fulfilling prophecies in the classroom to convey some of their strengths and weaknesses.

West and Anderson (1976) performed one of the first such studies. Based on a sample of 3,000 students, they found that a .12 effect of teacher expectations assessed in the freshman year of high school predicted sophomore achievement (controlling for freshman year achievement; [for the statistically inclined, all effects in naturalistic studies reviewed here are standardized regression coefficients, except where otherwise noted]). At about the same time, Williams (1976) analyzed teacher expectation effects among 10,000 high school students in Ontario high schools. There were eight possible self-fulfilling prophecy effects (two types of expectations were assessed by two different standardized test outcomes, computed separately for boys and girls). Despite the large sample size, seven of the eight were nonsignificant—in essence, zero. The eighth was .13.

Brattesani et al. (1984) examined the self-fulfilling effects of teacher expectations among 234 students in 16 fourth-, fifth-, and sixth grade classrooms. They found a slightly higher overall effect of about .26, for which there seems to be a likely explanation: They measured student initial achievement the prior year, teacher expectations in April of the current year, and student achievement at the end of the current year (presumably, May or June, though they did not specify the date). By April, after having had 8 months to observe their students, teachers would be likely to realize that some students were achieving at different levels than

indicated by their prior year standardized tests. If so, then accuracy likely contributes more substantially to their effect size than to effect sizes obtained in studies in which teacher expectations are assessed early in the current school year.

I also performed two studies of naturally occurring teacher expectation effects (Jussim, 1989; Jussim & Eccles, 1992). These studies included a total of about 100 teachers and 1,700 students in sixth grade math classes. Student initial achievement was assessed at the end of fifth grade and beginning of sixth grade; teacher expectations were assessed in October of sixth grade; and student standardized math test scores were assessed early in seventh grade. There were three teacher expectation variables, which we combined in all sorts of ways to predict changes in achievement. We stewed them; we fried them; we broiled them; and we braised them.³ But no matter what we did, the largest effects were only .18, and most were about .12.

More recently Trouilloud, Sarrazin, Martinek, and Guillet (2002) examined the self-fulfilling effects of seven swim instructors' expectations for 173 students' swimming ability. They assessed the extent to which swim instructors' expectations (assessed at the beginning of the 10-week class) predicted changes in how far students could swim in 10 minutes. Trouilloud et al. (2002) found a teacher expectation effect of .28, an effect slightly, but not dramatically, higher than the typically small self-fulfilling prophecies found in other studies.

These, however, are not the only naturalistic studies. Table 13-1 summarizes the results of every quantitative (i.e., they reported results in terms of numbers, not researchers' impressions) naturalistic study of self-fulfilling prophecies in the classroom of which I am aware. Two features of those results are particularly worth noting. Self-fulfilling prophecy effect sizes range from 0 to .4, with most falling between .10 and .20. Depending on how it is calculated, the overall mean effect size is between .07 and .17 (see Table 13-1 for more details). It remains unclear to me how researchers can justify testaments to the power of self-fulfilling prophecies (see quotes in Chapters 4 and 5) in the face of these data.

Second, Table 13-1 clearly shows that the larger the sample size, the smaller the self-fulfilling prophecy effect size (on average). Indeed, the correlation between sample size and effect size is $-.71$.

Why would larger samples produce such consistently smaller effects? This is the wrong question, because I am sure the large samples do not actually *cause* the smaller effects. Instead, as the statistically inclined can readily verify, effect sizes are more variable in smaller samples. Effect sizes that are near zero rarely get published. As the statistically inclined know, small sample studies *require* larger effect sizes than do large sample studies to obtain the holy grail of *statistical significance*. Without a statistically significant result, a self-fulfilling prophecy study is unlikely to get published. This primarily leaves the small-scale self-fulfilling prophecy studies that produce larger effects in the literature. The larger effect sizes obtained in the *published* small studies than in published large studies probably reflects the inherently greater random noise in such studies rather than any substantively generalizable evidence of larger self-fulfilling prophecies.

Expectancy-Confirming Biases in the Classroom

Whereas many studies have assessed self-fulfilling prophecies in the classroom, only a handful have addressed whether teachers' expectations bias their own judgments of students.

TABLE 13-1

Effect and Sample Sizes in Naturalistic Studies of the Self-Fulfilling Effects of Teacher Expectations		
Study	Self-Fulfilling Prophecy Effect Size	Sample Size
Williams (1976), ^a boys	.07	5,458
Williams (1976), ^a girls	.00	5,072
Chapman and McCauley (1993) ^b	.03	4,308
West and Anderson (1976) ^a	.12	3,000
Jussim and Eccles (1992) ^a	.13	1,288
Hinnant et al. (2009)	.11	693
Jussim (1989) ^a	.13	443
Doyle, Hancock, and Kifer (1972) ^b	.30	245
Brattesani et al. (1984) ^c	.26	234
Trouilloud et al. (2002)	.28	173
Kuklinski and Weinstein (2001), fifth grade ^a	.19	140
Kuklinski and Weinstein (2001), third grade ^a	.20	124
Kuklinski and Weinstein (2001), first grade ^a	.40	112
Palardy (1969) ^b	.14	107
Seaver (1973) ^b	.15	79

Note. The simple average of effect sizes, unweighted by sample size, is .17. The sample weighted average is .07. For this table, the correlation between sample size and self-fulfilling prophecy effect size is $-.71$. This means that studies with larger sample sizes generally found smaller effect sizes.

Williams (1976), Chapman and McCauley (1993), and Hinnant et al (2009) reported more than one self-fulfilling prophecy effect size. This table simply averaged them together. Hinnant et al (2009) had different sample sizes for each analysis; this table simply reports the average sample size. Williams (1976) performed analyses separately by student sex, and because these are two separate samples, they are treated as two studies. Kuklinski and Weinstein (2001) is treated as three separate studies because they performed analyses separately for first, third, and fifth graders. They actually reported two separate effect sizes for each grade, which, for simplicity, I have averaged together for this table.

^a Effect size reported as standardized regression coefficient.

^b These were quasi-experiments. Effect sizes are therefore reported as correlations between quasi-experimental conditions (reflecting teacher expectations) and student achievement.

^c Although this was a correlational study, path coefficients were not reported. Instead, they reported the r -squared increment obtained when adding teacher expectations to a model that included control variables. This table reports the square root of this value (also known as the semipartial correlation coefficient), which is more comparable to a correlation or regression coefficient.

Such biases themselves can be important because the grades students receive are based on teacher judgments. Furthermore, experimental research within social psychology has uncovered a host of errors and biases characterizing human judgment—this is so common that many social psychological reviews of social perception focus exclusively on error and bias without even mentioning accuracy (see Chapters 4 through 6 and 10). Such an emphasis would seem to predict that teacher expectations would lead to powerful biases in the classroom. Do they?

Rosenthal and Jacobson (1968a, b) and unexpected IQ spurts. The first study to at least partially address this issue was Rosenthal and Jacobson's (1968a, b) Pygmalion study, which found that teachers held very negative views of students who unexpectedly showed sharp IQ gains (see Chapter 3 for the details). With respect to understanding fundamental processes by which expectations influence judgments, such a result is important because it suggests people (teachers) prefer others (students) to behave (achieve) as expected. However, whether there were any concrete consequences of such bias remains unclear—after all, despite the teacher dislike, these students still showed dramatic IQ gains. Furthermore, whether their grades suffered was not evaluated.

Rosenthal and Jacobson (1968a, b) did, however, investigate whether there was any general tendency for teacher expectations to influence students' grades over and above the effect on IQ. They found no such evidence. The most direct test of whether teachers' expectations biased their judgments of students found no evidence of expectancy bias.

Rist (1970) and social class stereotypes. Rist's (1970) observational study (described in Chapter 4 and critiqued in Chapter 6) of a kindergarten class provided some of the earliest suggestion of teacher bias. Rist (1970) reported observing that teachers treated students from middle class backgrounds much more positively than they treated students from lower class backgrounds. He also presented some teacher quotes suggesting that, in contrast to the lower class students, the middle class students were functioning well in the classroom. Because Rist (1970) provided no quantitative data, however, it is impossible to determine the extent of such biases (and because this was an observational study in which Rist was the observer, it is impossible to determine the extent to which Rist's own hypotheses may have biased his observations or his reports).

Finn (1972) and bias in urban schools. Finn (1972) performed the first quantitative assessment of teacher bias. First, Finn had a group of fifth grade students write essays on topics such as "what I think about" and "describe your favorite subject." He then gave these essays to 300 fifth grade teachers in urban and suburban school districts to evaluate. Teacher expectations were manipulated by including (bogus) IQ information on the students who wrote the essays. Teachers in the high-expectation group were informed that the essay writer had an IQ in the 115 to 120 range; teachers in the low-expectation group were informed that the essay writer had an IQ in the 87 to 90 range.

Did high-expectation teachers evaluate the essays differently than did low-expectation teachers? No and yes. There was no expectancy effect in the suburban school districts—the essays were rated nearly identically, regardless of the teachers' beliefs about the students' IQ. In urban districts, however, the essays were rated more highly when the teachers believed they were written by high-IQ students (than when teachers believed they were written by low-IQ students). Even these effects were relatively small, however, ranging from about .1 to .2, depending on the analysis. Why the effects only occurred in urban districts was unclear, but Finn (1972) speculated that the higher quality of working and teaching conditions in suburban schools might help reduce bias.

Williams (1976): The clearest early report of teacher expectation bias in classrooms. Williams' (1976) study (described earlier in this chapter) found some of the clearest early evidence of teacher expectations biasing grades. Although Williams (1976) found that teacher expectations did not predict changes in students' standardized test scores, they did predict changes in students' grades—such effects ranged from .14 to .27. Higher expectancy students received

slightly higher grades than lower expectancy students, even when their achievement (as indicated by standardized achievement tests) was identical.

Jussim (1989) and Jussim and Eccles (1992): Bias in math classes. We assessed bias by examining the extent to which teacher expectations predicted changes in both grades and standardized test scores. Standardized test scores can be influenced by self-fulfilling prophecies, but not by teacher biases; grades can be influenced by both self-fulfilling prophecies and teacher biases. Therefore, we reasoned, the difference between the effects of teacher expectations on standardized tests and grades constituted a way to assess bias. If teacher expectation effects on grades were similar to those of standardized test scores, self-fulfilling prophecy would provide a simple and sufficient explanation for both. If, however, teacher expectation effects on grades were larger than those on standardized test scores, it would suggest that both self-fulfilling prophecies and biases influenced grades.

Consistent with Williams' (1976) research, teacher expectations predicted changes in grades more strongly than they predicted changes in standardized test scores. Effects on grades generally ranged from about .2 to about .4; effects on standardized tests ranged from about .1 to .2. Thus, bias seemed to increase the teacher expectation effect on grades about .1 to .2 over the effect of self-fulfilling prophecy.

Trouilloud et al. (2002): No bias in swim classes. Like Rosenthal and Jacobson's (1968a, b) study, the Trouilloud et al. (2002) study of swim classes (described earlier) looked but found no evidence of teacher expectations biasing students' grades. In fact, they found that teacher expectations predicted achievement (swimming distance in 10 minutes) more strongly than grades, a result completely inconsistent with the bias hypothesis. Overall, therefore, it is clear that biases, like self-fulfilling prophecies themselves, are generally quite small and far from inevitable.

The Limited Power of Teacher Expectations: Conclusions

Social psychological claims about the power of expectancy effects notwithstanding (see Chapters 4, 5, and 6), *teacher* expectation effects are typically quite small. The self-fulfilling effects of teacher expectations have now been assessed in a wide variety of experimental and naturalistic studies conducted at virtually every grade level from kindergarten through 12th, in several different countries, and in several different types of classes (English, reading, literature, math, and swimming). This is a wide variety of evidence collected by a wide variety of researchers and conducted in a wide variety of contexts. As such, it constitutes a very strong empirical, scientific base for reaching general conclusions about the typical power of teacher expectation effects. And it is vividly clear that such effects are, on average, around .1 to .2.

Is this reasonably described as "large," "powerful," or "dramatic"? This is subjective for two reasons: (1) Reasonable people may differ on the size of an effect considered to be dramatic, and (2) words like "powerful" ("large," "dramatic," etc.) are inherently ambiguous, so that two different people might mean something different and use the same word to describe it. Does this mean there is no way to evaluate whether it is reasonable to characterize the typical self-fulfilling prophecies in the classroom as powerful and dramatic?

Well, one definition of dramatic is "surprising." The discovery of life, even very simple life, on Mars would certainly be dramatic. In a similar (but perhaps somewhat less dramatic)

spirit, that teachers' expectations ever create self-fulfilling prophecies could reasonably be described as a "dramatic" finding.

How about "powerful"? The thing about slippery concepts like this is that we need some sort of a priori standard—so that we can't go around calling effects we like as "powerful" and effects we do not like as "trivial." A long-standing convention within psychology has been to characterize effects of above .4 as large, those between .2 and .4 as moderate, and those below .2 as small (Cohen, 1988). Of course, just because Cohen (1988) said such effects are small does not mean everyone else has to agree with him in any particular case. Nonetheless, by this standard, it is clear that self-fulfilling prophecy and biasing effects of teacher expectations are typically small and only rarely even reach "moderate."

So, let's consider other ways to evaluate whether it is reasonable to consider teacher expectation effects as "powerful." One such way would be empirical—to compare teacher expectation effects to effect sizes typically obtained in psychological research. By this criteria, the average teacher expectation effect size falls in the bottom third of effect sizes obtained in 380 meta-analyses (Hemphill, 2003) and in the bottom half of the effect sizes in all of social psychology (Richard, Bonds, & Stokes-Zoota, 2003).

So far, both an a priori determination of what constitutes a powerful effect size and an empirical assessment of what constitutes a powerful effect size both lead to the conclusion that teacher expectations are not very powerful. But these are both very abstract criteria—wouldn't it be nice if there was a more concrete, real-world way to see if the effect is powerful? It turns out that there is.

Specifically, we can translate the average expectancy effect size into a metric that indicates the proportion of students in a particular class likely to be affected by self-fulfilling prophecies. Over 20 years ago, Brophy's (1983) narrative review concluded that, on average, teacher expectations typically have self-fulfilling effects on only 5% to 10% of students. This conclusion has held up remarkably well. As shown in Table 13-2, Rosenthal's (1984) binomial effect size display (BESD) also shows that the typical teacher expectation effect of .1 to .2 means that self-fulfilling prophecies typically change the achievement of about 5% to 10% of all students.

Is 5% to 10% large? I once wrote that it could be considered large (Jussim, 1990), because an intervention that substantially increased the achievement of 10% of all students would be hailed as a major policy success. If there are a billion schoolchildren around the world, and self-fulfilling prophecies affect 5% to 10% of them, we are talking 50 million to 100 million people. Not too shabby.

But I have come to have more reservations about this view than when I first wrote about it. First, I have come to view the 50 million to 100 million number as little more than intellectual gymnastics. This is because even the tiniest, most trivial effect, if multiplied by a sufficiently large number of instances or people, will produce a large number. This does not make the effect large.

Another way to put more of a realistic context on evaluating the size of the effect is to bring it down to earth—how many students in a typical American classroom are likely to be affected by self-fulfilling prophecies? Changing the performance of 5% to 10% of the students in a class means changing the performance of 1 or 2, in a class of 20. This is the same as saying self-fulfilling prophecies *do not* affect 18 to 19 students in our class of 20.

Or, consider another hypothetical but concrete example. Consider a superintendent of a school district who has come under fire because half the students in his district failed a state-wide standardized test. Our hypothetical superintendent then calls a press conference and

TABLE 13-2

Correlations of Teacher Expectations with Student Achievement and Their Self-Fulfilling Effects

Teacher Expectations Have No Effect on Student Achievement:

	Low Teacher Expectations	High Teacher Expectations
Students with above-average future achievement	50%	50%
Students with below-average future achievement	50%	50%

Regardless of whether teacher expectations are high or low, 50% of students end up with above-average achievement and 50% end up with below-average achievement.

Teacher Expectations Have an $r = .1$ Self-Fulfilling Effect on Student Achievement:

	Low Teacher Expectations	High Teacher Expectations
Students with above-average future achievement	45%	55%
Students with below-average future achievement	55%	45%

Fifty-five percent of high-expectation students perform above average, whereas only 45% of low-expectation students perform above average. High expectations increase the performance of 5% of the students and low expectations decrease the performance of 5% of the students.

Teacher Expectations Have an $r = .2$ Self-Fulfilling Effect on Student Achievement:

	Low Teacher Expectations	High Teacher Expectations
Students with above-average future achievement	40%	60%
Students with below-average future achievement	60%	40%

Sixty percent of high-expectation students perform above average, whereas only 40% of low-expectation students perform above average. High expectations increase the performance of 10% of the students and low expectations decrease the performance of 10% of the students.

declares, "Our failure rate is unacceptable. I have a plan that will produce powerful and dramatic increases in student achievement over the next few years." Over the next few years, instead of 50% failing, 42% fail. Do you think most people would think he made good on his promise to produce powerful and dramatic increases in achievement? Regardless of your answer here, I think the wealth of accumulated data on teacher expectations justifies generalizations that emphasize their limited power, rather than those that characterize such effects as unusually influential.

Less research has addressed the bias question. Nonetheless, evidence that teacher expectations bias their perceptions and evaluations of students has consistently emerged from observational, naturalistic, and experimental studies. Such effects, too, are typically relatively small—averaging about .2, as indicated by the studies reporting results that could be quantified.

Nonetheless, the *total* expectancy effect on grades, including both self-fulfilling prophecy and bias, may not be quite so small. A small self-fulfilling prophecy of .1 to .2 plus a small bias

of about .2 will yield a total expectancy effect of .3 to .4. The effects may add up to something substantial enough to make a noticeable difference among some students.

To get concrete, if some high-expectancy students benefit from a self-fulfilling prophecy such that their objective achievement increases from meriting a B to a B+, and these students benefit from a teacher bias on top of the self-fulfilling prophecy, those students may end up with As. Thus, somewhat better than average students have attained a record of excellence—a pattern that may at least partially explain the apparent academic “success” of teachers’ pets. Similarly, if the same B students were victims of a low teacher expectation, self-fulfilling prophecy might lead their performance to decline to a C+. If further victimized by a negative bias effect, they may end up with Cs in the class. Thus, somewhat better than average students have attained a fairly poor record in this class. When taken together, if both bias and self-fulfilling prophecy occur at the same time, in the same direction, for the same target (or student), their combined effect can be substantial.

Why Are Teacher Expectations Effects So Weak?

The conclusion that the combination of self-fulfilling prophecy and bias *can sometimes* lead to a substantial effect on grades should not be (mis)interpreted as meaning that such effects are common or typical. *Both* self-fulfilling prophecies and expectancy-confirming grading biases in the classroom are generally sufficiently small that, in general, only a small minority of students will be affected by either one (see Table 13–2). Even fewer will be substantially affected by both. Especially within the context of the traditional emphasis within social psychology on the power of such effects, and the extraordinary power differential between teachers and their innocent and nearly defenseless elementary school students, it is natural to wonder why such effects are typically so weak. What prevents students from readily caving in to teachers who, consciously or not, seek to impose their expectations?

Most of the answer to this question did not come from social psychology for two reasons: (1) Because of the traditional social psychology emphasis on the power of expectancy effects, the question was slow to arise, and (2) accuracy is one of the most likely explanations for weak expectancy effects (accurate expectations do not produce self-fulfilling prophecies) and accuracy research was banished from social psychology for about 30 years. It should not be surprising, therefore, that social psychology has produced relatively little research addressing the accuracy of teacher expectations (other than my own). Classic social psychology, however, analyzes many phenomena from several different perspectives: the perceiver, the target, their interaction, and their situation. Let’s see how far we can get toward understanding the weakness of teacher expectation effects using these classic social psychological tools.

Teachers: The Accuracy of Their Expectations

Only inaccurate expectations can produce self-fulfilling prophecies; accurate expectations do not (see Chapters 2 through 4). Therefore, the more accurate expectations teachers develop, the less potential there is for self-fulfilling prophecy. Teacher accuracy, therefore, may be one explanation for typically weak expectancy effects.

Teacher expectations as predictors of student achievement. How accurate are teacher expectations? Simple correlations between teacher expectations and student achievement are often moderately to very high, typically ranging from about .4 to about .7 (Brattesani et al., 1984; Brophy & Good, 1974; Crano & Mellon, 1978; Hoge & Butcher, 1984; Humphreys & Stubbs, 1978; Jussim, 1989; Jussim & Eccles, 1992; Williams, 1976). Such correlations mean that teacher expectations are typically quite good predictors of students' future achievement. A correlation of .5 means, for example, that 75% of the students teachers identify as being top achievers early in the year end up as top achievers at the end of the year.

How to interpret these correlations, however, is less clear. Do they reflect accuracy, self-fulfilling prophecy, or some combination of both? The next sections discuss how to answer these questions.

Inaccurate but uninfluential. "What about the other 25%?" you may be wondering. "Isn't that degree of inaccuracy ample room for self-fulfilling prophecy?" This suggestion may sound reasonable until one really thinks about it. By definition, if 25% of the students *did not* confirm the teacher's expectation, we know not only that those expectations were inaccurate but also that they did *not* cause a self-fulfilling prophecy. How? If they had, those students would have confirmed the teacher's expectation! Thus, teacher expectations were indeed inaccurate for these students but, despite their inaccuracy, produced no self-fulfilling prophecies. Thus, the only students among whom the relative roles of accuracy and self-fulfilling prophecy can even be compared are the 75% who confirmed the teacher's expectations.

The causal ambiguity of the correlations. On their own, however, simple correlations between teacher expectations and student achievement are interpretively ambiguous because they represent some unknown combination of accuracy and self-fulfilling prophecy (when the outcomes are standardized tests) and also expectancy-confirming bias (when the outcomes are grades). Thus, one cannot conclude *either* that self-fulfilling prophecies or accuracy constitutes the main explanation for students confirming teacher expectations simply on the basis of these correlations.

Assessing accuracy. In contrast to social psychology, within educational psychology, assessing the accuracy of teacher expectations was never viewed as unusually problematic and was accomplished in two main ways. First, the results of studies that simply correlated teacher expectations with student achievement (e.g., Brophy & Good, 1974) were compared with the self-fulfilling effects of teacher expectations obtained in experimental studies. The logic here is quite straightforward: (1) Correlations represent the overall extent to which teacher expectations predict student achievement; (2) that overall predictive power derives from some unknown combination of accuracy and self-fulfilling prophecy; (3) experiments, which are supremely well suited for identifying causality, provide information about the causal effects of teacher expectations; and (4) the overall predictive validity minus the self-fulfilling prophecy effect equals the extent to which teacher expectations predict but do not cause student achievement—that is, accuracy.

This line of research consistently showed that the correlations between teacher expectations and student achievement were typically much higher (generally in the .4 to .7 range) than were the expectancy effect sizes (.1 to .2 range—see, e.g., Brophy, 1983; Jussim, 1991, for reviews). By this metric, about 75% of the overall predictive validity of teacher expectations for standardized test scores reflects accuracy and the remaining 25% reflects self-fulfilling

prophecy. Of course, such evidence is only indirect because it involves comparisons of results across different, and often disparate, studies, rather than demonstrating a 75% accuracy/25% self-fulfilling prophecy pattern within a study.

It was, therefore, important to test the relative degrees of accuracy and self-fulfilling prophecy within a single study. Doing so required (1) assessing teacher expectations (typically early in the school year); (2) assessing student achievement in the year prior to the assessment of teacher expectations; (3) assessing student outcomes at some later point—typically, the end of the school year in which teacher expectations were assessed; and (4) examining the extent to which teacher expectations early in the year predicted (but did not cause) student outcomes late in the year (which was accomplished by controlling for student performance prior to the assessment of teacher expectations).

Such tests yielded conclusions essentially identical to those obtained through indirect comparisons of results across different studies. Table 13–3 summarizes the key results obtained in almost every naturalistic study of teacher expectations I could find that was capable of simultaneously examining the extent of self-fulfilling prophecy and accuracy. The first three columns of results in Table 13–3 are relatively simple and straightforward. The first just identifies a particular study. The second column presents the simple correlations between teacher expectations and student achievement. Those correlations range from moderate (.36) to quite high (.79) and mean that teacher expectations typically predict student achievement at least reasonably well and sometimes extraordinarily well.

TABLE 13–3

Relations Between Teacher Expectations and Student Achievement: Correlations, Self-Fulfilling Prophecies, and Accuracy			
<i>Study</i>	Correlations of Teacher Expectations with Student Achievement	Self-Fulfilling Prophecy Effects	Correlation Minus Self-Fulfilling Prophecy Equals Accuracy
Williams (1976)	.47–.72	.00–.13	.42–.72
Brattesani et al. (1984)	.74	.26	.48
Jussim (1989)	.36–.57	–.03–.18	.36–.41
Jussim and Eccles (1992)	.50–.55	.10–.16	.36–.49
Kuklinski and Weinstein (2001)	.50–.70	.03–.40	.10–.54
Trouilloud et al. (2002)	.79	.28	.51

Self-fulfilling prophecy effects are standardized regression coefficients relating teacher expectations earlier in the year to student achievement later in the year. All such effects were obtained in the context of models that controlled for students' prior achievement (often, there were other controls, too—see the original studies for the details). When a range for the correlation, self-fulfilling prophecy, or accuracy is presented, it is because these studies measured more than one type of teacher expectation and/or examined relations with more than one type of achievement, thereby yielding multiple correlations and effects.

The question is, Why do teacher expectations predict student achievement so well? One possibility is that teacher expectations create large and powerful self-fulfilling prophecies; another is that they are accurate. Table 13–3 shows how to separate out, and empirically test, the extent to which self-fulfilling prophecies versus accuracy account for the extent to which teacher expectations predict student achievement.

The standardized path coefficients (the entries in Table 13–3 in the column labeled “Self-Fulfilling Prophecy Effects”) link teacher expectations to student achievement in the context of a model that controls for plausible sources of accuracy (student prior grades and achievement, demographics, motivation, etc.). These coefficients represent the best estimate of the extent to which teacher expectations early in the year predict *changes* in student achievement by the end of the school year (we know this because prior achievement is controlled). The difference between the overall predictive validity of teacher expectations (the correlation with achievement) and the standardized path coefficient estimating self-fulfilling prophecy equals the extent to which teacher expectations predicted but did not cause student achievement. Prediction without causation is accuracy.³

Simply eyeballing the results in the first two columns of Table 13–3 shows quite clearly that the simple correlations are generally much larger than the expectancy effects. This alone should produce a sort of “aha!” reaction among those of you with at least enough statistical training to understand correlations. Clearly, teacher expectations predict student achievement far more successfully than can be accounted for merely by self-fulfilling prophecies. The extent to which they do so is accuracy.

The third column of data, the accuracy column, shows precisely how much accuracy contributes to the overall correlation. Without getting mathematically heavy here, the logic is quite straightforward:

1. The correlation of teacher expectations with student achievement is the overall predictive validity.
2. Each study then assessed the self-fulfilling effects of teacher expectations.
3. Predictive accuracy, which I define as predictive validity without self-fulfilling or causal influence, then simply equals the difference between the correlation and the self-fulfilling prophecy effect. As shown in Table 13–3, accuracy typically accounts for 65% to 100% of the correlation between teacher expectations and student achievement, whereas self-fulfilling prophecy only accounts for 0% to 35%. In short, although self-fulfilling prophecies clearly occur, (a) they tend to be small and (b) teacher expectations predict student achievement primarily because they are accurate.

Of course, naturalistic studies are not experiments. Causal conclusions reached on the basis of such studies must be considerably more cautious than those based on experiments. This, however, constitutes a threat not to the accuracy interpretation of such studies, but to the self-fulfilling prophecy interpretation! No matter how well conducted any naturalistic study is, it is always possible that it has omitted some important third variable. If any study summarized in Table 13–3 omitted an important variable that predicted both teacher expectations and student achievement late in the year, that study overestimated self-fulfilling prophecies and underestimated accuracy. Such a variable, if it exists, would mean that teacher

expectations are *even less powerful* than I have concluded, and that *accuracy is even higher than I have concluded*. Thus, one could consider the accuracy estimates in the final column of Table 13–3 to be **lower bounds** on the likely degree of accuracy found in that study.

The bottom line, however, has been that studies using this approach yielded essentially the same results as the cross-study comparisons (see reviews by Brophy, 1983; Jussim & Eccles, 1995). Although self-fulfilling prophecies do occur, teacher expectations predict student achievement mainly because those expectations are accurate.

Students: Why It May Not Be So Easy to Change Them

Intelligence, skill, competence. The malleability of intelligence has long been a subject of hot scholarly debate, with some folks arguing or implying that, after about age 10 or so, intelligence remains largely unchanged throughout one's lifetime, and others suggesting that intelligence can be changed through education, training, and experience (e.g., Herrnstein & Murray, 1994; Neisser et al., 1996). Although resolving this debate is beyond the scope of this chapter, one thing is clear: It is, at best, extraordinarily difficult to change intelligence. Programs designed to do so typically produce short-term changes that equally typically do not last long after the programs end (e.g., Neisser et al., 1996).

This should not be surprising. Developing skills and competencies at almost anything—serving a tennis ball, preparing an elegant meal, or creating a work of art—takes a long time. One does not first learn how to avoid burning the French toast today and tomorrow become capable of personally preparing a seven-course meal complete with hors d'oeuvres, bisque, Mesclun salad, pheasant under glass, and a chocolate fondue for dessert. Even if intelligence is malleable, surely it is no more malleable than cooking skill. And developing skill at pretty much anything takes time, effort, and experience.

Furthermore, once a high level of skill (at almost anything) is attained, it will typically decline without continued practice and effort. This is painfully salient to me. I suffered a rash of injuries from 1997 to 2005 that kept me off the tennis court for long stretches. Each time, when I finally got back to playing, I not only had to overcome whatever pain or residual injury I still had, but I also needed to overcome the extraordinary extent to which the quality of my play had deteriorated from mere lack of play.

Thus, changing skill levels, whether reading skill, math skill, athletic skill, cooking skill, or general intelligence, clearly is difficult and time-consuming. Furthermore, without relentless and continued practice, the skills tend to dissipate. When the extraordinary difficulty in developing and maintaining skill at almost anything is considered, it is perhaps considerably less surprising that self-fulfilling prophecy effects in the classroom tend not to be particularly powerful. Skill, and especially intellectual skills, may be movable, but they move very slowly. They are just not that easy for anything, including others' expectations, to push around.

Self-verification. Chapter 7 discussed at length the potential power of self-verification to limit self-fulfilling prophecies. In short, students typically know quite a lot about who they are and what they are good and not good at. When others, including teachers, express manifestly inaccurate views about them, many students will not readily cave in. Instead, students may frequently resist that expectation and work hard to convince the teacher to view them much as they view themselves. When self-verification succeeds, as it often does (see Chapter 7),

teachers will end up changing their expectations, rather than imposing their expectations on students. The only study to address this process in an educational context showed that students self-verify (convince teachers to view them much as they see themselves) to about the same extent that teachers' expectations influence student self-concepts—and both effects were quite small—around .1 (Madon et al., 2001).

Teacher–Student Relationships

Teachers are, obviously, much more powerful than students, and people in high-power positions are more capable of imposing their expectations on people in low-power positions. However, the major phenomenon to be explained is not why teacher expectations are so powerful; it is why they typically are so weak. Furthermore, the power differential between teacher and students is more or less constant, so that it seems unlikely to explain why self-fulfilling prophecies are usually weak but occasionally strong.

One likely explanation for the typically weak self-fulfilling effects of teacher expectations is the length of the teacher–student relationship. Most experimental studies of expectancy effects focus on initial interactions between strangers. Although expectancy effects are not particularly large even in those contexts (see Chapters 6 through 9), they are likely to be even smaller in long-term relationships for several reasons.

First, in general, interpersonal expectancies are more likely to be inaccurate in initial interactions between strangers than in interactions among folks who have known each other a long time. The longer a relationship lasts, the more information each person has about the other, so the less chance there is for dramatic inaccuracies to be maintained. Length of relationship does not guarantee perfect accuracy, and some degree of error, bias, and imperfection may indeed be maintained even in long-term relationships. Nonetheless, evidence from research on stereotyping, self-verification, and teacher expectations (e.g., Eagly, Makhijani, Ashmore, & Longo, 1991; Krueger & Rothbart, 1988; Raudenbush, 1984; Swann & Ely, 1984) supports the following conclusions: (1) The more personal information perceivers have about targets, the less they rely on stereotypes when judging those targets; (2) given multiple opportunities for interaction, perceivers are more likely to alter their erroneous expectations to fit the target than to alter the target to fit their expectations; and (3) the longer teachers know their students, the less likely it is for inaccurate information to influence their expectations. Because of the frequency with which student performance is evaluated in the classroom, teachers have many opportunities to obtain information indicating a need to alter their expectations. All but the most oblivious or rigid of teachers will likely alter their beliefs about particular students whose performance is manifestly different than the teacher first expected.

“But,” you may be wondering, “what happens when a student does get one of those rigid or oblivious teachers?” Good question. There is good evidence that self-fulfilling prophecies are stronger among such people (e.g., Babad, Inbar, & Rosenthal, 1982; Brophy, 1983; Harris & Rosenthal, 1985). However, even here the length of teacher–student relationships typically builds in a natural limit to expectancy effects. Specifically, most classes are taught by a teacher for one school year. This builds in a natural limit to the self-fulfilling effects of even an evil, oblivious, rigid teacher (unless one assumes that the next teacher is likely to be equally

evil, oblivious, rigid, and inaccurate—and the general issue of whether expectancy effects accumulate or dissipate over time will be addressed in the next chapter). Thus, even though the teacher–student power differential probably increases the potential for self-fulfilling prophecies, this potential is probably more than counterbalanced by the combination of lots of performance information plus long-term relationship, which functions primarily to reduce and limit the extent of self-fulfilling prophecies in the classroom (see also Brophy & Good, 1974, for an extended discussion of teacher–student relationships with a particular emphasis on understanding self-fulfilling prophecies).

Students' Situations

There is a whole world of stuff outside of school, good or bad, that profoundly influences student achievement, regardless of what teachers do in their classes. Physical and mental health (of both students and their immediate family), social class, cultural background, and family emphasis on academics are all factors that (1) have major effects on students' learning, motivation, and achievement and (2) are not likely to be greatly affected by teachers' expectations. Students who have a particular academic trajectory (good or bad) produced by these other factors are not likely to have their course greatly altered by a single teacher's erroneous expectations. Aspects of students' situations outside of school may often serve as an anchor, making it difficult for teachers to dramatically alter the achievement of their students.

Conclusions: Why Classroom Expectancy Effects Are Typically Small

When all the factors operating *against* expectancy effects in the classroom are thoughtfully considered, it is a wonder that they occur at all, not that they are typically small. When faced with clear, objective information about another person, people, including (but not restricted to) teachers, typically use that information in making judgments. Most people hold most of their interpersonal expectations rather flexibly and readily change them when faced with disconfirming information. Thus, one reason inaccurate expectations may not be self-fulfilling very often is that those expectations are changed when people are faced with new information. In short, people are often reasonably accurate when they have useful information about others.

Another reason inaccurate expectations are not necessarily self-fulfilling is that students are not usually passive receptacles waiting to be filled by teacher beliefs. They have their own views of themselves and, when they believe the teacher's view is incorrect, may often resist confirming it, and even attempt to change that view. Last, students' situations outside of school, for better or worse, often have substantial influences on their motivation and resources that constrain the ability of teachers to dramatically alter students' academic trajectories.

None of this, however, denies the *possibility* that self-fulfilling prophecies in the classroom can sometimes be quite powerful. Indeed, the next section reviews attempts to identify those conditions under which truly powerful self-fulfilling prophecies occur. Unfortunately, however, there has been far less research on this issue than on other aspects of teacher expectations. This may be an unintended negative side effect of the long-standing overemphasis on the power of such effects. If researchers erroneously believe powerful self-fulfilling

prophecies have routinely been found, there would seem to be little need to search for such effects.

Ironically, however, powerful effects have almost never been found among the research most frequently cited as testaments to the power of expectancy effects (see Chapters 6 and 9). This raises the question: Speculation aside, has empirical, scientific research ever found powerful self-fulfilling prophecy effects? To answer this question, the next section reviews the relatively sparse research that has sought to identify conditions under which bona fide powerful self-fulfilling prophecies occur in the classroom.

Self-Fulfilling Prophecies Among the Downtrodden: The Core Attention-Grabbing Value of Pygmalion

Prior chapters have reviewed the ways in which Rosenthal and Jacobson's (1968a,b) Pygmalion study (Chapter 3) and the subsequent social psychological research on the self-fulfilling nature of social stereotypes (Chapter 4) ignited the interest of social scientists for decades. One such reason, and perhaps the main one, was that self-fulfilling prophecies seemed to provide a relatively benevolent social forces explanation for the underachievement of students from stigmatized or disadvantaged backgrounds, such as racial and ethnic minorities, students from lower social class backgrounds, etc.⁴ Indeed, the short report of Pygmalion that appeared in *American Scientist* (Rosenthal & Jacobson, 1968b) was titled "Teacher Expectations for the Disadvantaged" precisely because of the presumed implications regarding sources of, and ways to raise, the achievement of students from disadvantaged backgrounds.

Thus, even if self-fulfilling prophecies are not typically large, a large part of their appeal might be salvaged if it turned out that expectancy effects were much larger among disadvantaged students than among other students. Might self-fulfilling prophecies be unusually powerful among, for example, African Americans and students from lower social class backgrounds? It remains amazing to me that so little empirical research, either experimental or naturalistic, has actually examined whether such groups are more vulnerable to expectancy effects. Given the frequency with which researchers point to self-fulfilling prophecies as a source of social inequalities (e.g., Darley & Fazio, 1980; Gilbert, 1995; Jones, 1986; Weinstein et al., 2004), this remains a stunning gap in the scientific investigation into relations between social beliefs, social reality, and social problems. Nonetheless, one program of research did investigate these issues a few years ago—and the next section describes what we found out.

Our "Quest" for the Powerful Self-Fulfilling Prophecy

Having discovered over a period of nearly 10 years that (1) my own studies of expectancies consistently found only modest self-fulfilling prophecy effects and (2) most other studies actually found something quite similar, it seemed to me important to try to discover if there were *any* conditions under which truly powerful self-fulfilling prophecies in the classroom occurred. Thus, starting about 1994 (first paper published in 1996—Jussim et al., 1996), we (primarily Jacquelynne Eccles, Stephanie Madon, Alison Smith, and myself, but also several

other graduate and undergraduate students) embarked on a quest to systematically search for conditions under which large expectancy effects occurred.

The data, the model, and the analyses. All “quest” studies described in this chapter were based on the Michigan Study of Adolescent Life Transitions (MSALT), which assessed a variety of social, psychological, demographic, and achievement-related variables in a sample that included over 200 teachers and 2,000 students in sixth and seventh grades (see Eccles et al., 1989; Midgley, Feldlaufer, & Eccles, 1989; Wigfield, Eccles, MacIver, Reuman, & Midgley, 1991, for more details about the data).⁵

The quest studies were not experiments—they were entirely based on real-world, naturalistic (correlational) data. Therefore, to reduce as much as possible non-self-fulfilling prophecy interpretations of our results, teacher expectations were assessed early in the school year (October) and student achievement was assessed at the end of the school year (final grades) or early the following year (standardized test scores). The longitudinal (over time) nature of the data means that we can be certain end-of-year achievement did not cause early-year teacher expectations (the future cannot possibly cause the past). Furthermore, all analyses controlled for students’ prior year grades and scores on standardized tests taken prior to the assessment of teacher expectations. These controls set a high hurdle for early-year teacher expectations: They could not merely predict student achievement—to be interpretable as probable evidence of a self-fulfilling prophecy, they had to predict *changes* in future student achievement.⁶

Self-fulfilling prophecies among students subject to stigma. Several theoretical arguments led us to suspect that students from stigmatized groups would be more susceptible to self-fulfilling prophecies than are other students. Abundant evidence suggests that school is often an unfriendly place for many African American and lower socioeconomic status (SES) students (e.g., Lareau, 1987; Steele, 1992). When school is consistently a difficult place, students may often “disidentify” with achievement by devaluing the importance they place on school or by devaluing the particular subjects in which they feel devalued (e.g., Eccles (Parsons), 1984; Eccles et al., 1983; Eccles & Harold, 1992; Jussim, 1986; Meece, Eccles-Parsons, Kaczala, Goff, & Futterman, 1982; Steele, 1992). Such responses may render them more readily influenced by teacher expectations in several ways.⁷

When students with a history of negative school experiences find themselves faced with a supportive, encouraging teacher who also insists on high performance, it may feel like a breath of fresh air. Such a teacher may inspire previously low achievers to new heights. This perspective may not be as Pollyanish as it sounds. In his influential article on Black disidentification with school, Steele (1992) describes academic programs in which previously low-performing students (e.g., some with SATs in the 300s) take on difficult honors-level work and come to outperform their White and Asian classmates. Steele’s (1992) description of these programs implied that the teachers often engage in behaviors much like those that lead to beneficial self-fulfilling prophecies in the classroom and workplace: They are challenging and supportive (e.g., Brophy & Good, 1974; Cooper, 1979; Eccles & Wigfield, 1985; Eden, 1984, 1986; Harris & Rosenthal, 1985; Jussim, 1986; Rosenthal, 1989). With these ideas in mind, we examined whether students stigmatized by race, social class, or their own low achievement were more susceptible to self-fulfilling prophecies.

Self-fulfilling prophecies in Black and White. This section title refers not merely to the fact that we studied whether race/ethnicity moderated self-fulfilling prophecies—it nicely

captures the stark and striking difference in the power of self-fulfilling prophecies among White and African American students that we ultimately found. Specifically, the self-fulfilling effect of teacher expectations on African American students' achievement was well over double the size of the self-fulfilling effect of teacher expectations on White students' standardized test scores and grades.⁸ For White students, our results indicated that being the target of the lowest versus highest teacher expectations could make as much as a 20-percentile-point standardized test difference, although the typical difference was more like 5 to 10 percentile points.

These students took a standardized statewide test called the Michigan Educational Assessment Program in sixth grade. To put this pattern on a scale more familiar to most people, I have translated these patterns into comparable SAT scores (but please keep in mind that these students were in sixth grade and this is just an analogy and an approximation to help make the meaning of these differences more concrete). The results for White students were comparable to, at most, going from about 480 to about 520 on the SATs and, more typically, going from about 490 to about 510. So, for White students, it was definitely better to have the teacher think highly of them—but unless the difference was between teachers thinking a student was almost incompetent versus absolutely brilliant, teacher expectations only made a modest difference.

Among African American students, however, a very different picture emerged. *On average*, teacher expectations made a nearly 20-percentile-point standardized test score difference. And being the target of the lowest versus highest teacher expectations could make as much as a 58-percentile-point standardized test score difference. This means that, in the extreme, the difference between being the target of negative versus positive expectations could have made a difference as large as going from about 420 to 580 on the SATs, and even a typical difference would have been similar to going from about 480 to 520.⁹

Social class and self-fulfilling prophecies. Were students from lower SES backgrounds also more vulnerable to self-fulfilling prophecies? They were.¹⁰ Teacher expectations did not predict changes in the achievement of middle class students. Among lower class students, however, teacher expectations, on average, made a 10- to 15-percentile-point difference. And being the target of the lowest versus highest teacher expectations could make a difference of over 40 percentile points. By analogy with SATs, these differences correspond roughly to the difference between 490 and 520 for a typical effect and the difference between 450 and 560 for the largest possible effect.

Previous achievement and social class. We also suspected that students with histories of low achievement might be particularly vulnerable to teacher expectations. Students seem most likely to disidentify with school when school becomes a painful place (either because of failure or cultural devaluation—see Steele, 1992). Disidentification means, in part, investing less energy in schoolwork, thereby leading to lower academic performance. These negative past experiences may also create (1) a readiness to tune out of school as soon as it becomes troublesome yet again (e.g., in a new school year) and (2) a heightened receptiveness to teachers who treat them with care and respect while at the same requiring them to meet high standards.

Our initial analyses (Madon, Jussim, & Eccles, 1997) addressing this issue with the MSALT data found some support for this perspective: Teacher expectations did indeed predict the future achievement of students with histories of prior low achievement more strongly than

they predicted the future achievement of students with histories of high achievement.¹¹ Among high achievers, the typical effect was comparable to going from about 490 to 510 on the SAT and the largest effect was comparable to going from about 470 to 530. Among low achievers, however, the typical self-fulfilling prophecy effect was comparable to going from an SAT score of 480 to 510 and the largest effect was comparable to going from 450 to 550.

Not bad, but we (Madon et al., 1997) did not stop there; instead, we continued to explore this possibility from other, related angles. First, we speculated that the effects of teacher expectations on student achievement might not be linear (e.g., high expectations might lift student achievement more than low expectations reduce student achievement). We examined this question by (1) first determining whether teacher expectations were higher or lower than they should be, based on student prior achievement, and then (2) using these high or low teacher expectations to predict changes in student achievement over the school year. And, indeed, we found some modest evidence that positive expectancies improved student achievement more so than negative expectancies harmed achievement.

Up to this point we had discovered quite a few conditions under which self-fulfilling prophecies were larger than usual. Next, we wondered whether these conditions might combine to produce some quite large effects among at least some students. Although there were too few African American students for us to examine ever-smaller subsamples, there were more than enough targets of high expectations, low achievers, and students from lower socioeconomic backgrounds.

First, we found that high expectations had uniquely powerful self-fulfilling effects on low achievers. Whereas both high and low teacher expectations had either no or modest effects on high achievers, and whereas low teacher expectations also had only small effects on low achievers, high teacher expectations predicted dramatically higher achievement among low achievers.¹² The self-fulfilling prophecy effects among the low-achievement/high-teacher-expectation students were about the same magnitude as the effects among African American students.

We (Jussim, Eccles, & Madon, 1996) also examined whether low achievement might combine with social class to produce a uniquely powerful vulnerability to self-fulfilling prophecies. The self-fulfilling prophecy effects among most combinations of social class and achievement were similar to those found in our analyses focusing just on social class. However, the effects among students from lower social class backgrounds and with histories of low achievement were the strongest found in our quest studies, and among the strongest ever found in self-fulfilling prophecy research.¹³ They produced effects on standardized achievement tests roughly comparable to the difference between 470 and 530 on the SATs for a typical effect, and for differences comparable to going from under 400 to over 600 for the largest possible effects. These were dramatic differences.

Limitations to Our “Quest” to Discover Powerful Self-Fulfilling Prophecies

The biggest limitation is that there have been few attempts to replicate this research. Although we used relatively large samples for most of our analyses, all studies reported here are just that—single studies. And I would almost never recommend treating a conclusion as “fact” based on a single study, even my own. Whether these patterns are common and general or were somehow unique to this particular data remains a question for future research.

Nonetheless, those patterns do at least raise the possibility that relatively powerful expectancy effects do systematically occur under conditions that correspond well with the traditional social psychological emphasis on social issues.

One recent study did address several of the “quest” issues among younger students. Hinnant, O’Brien, & Ghazarian (2009) assessed relations of first grade teachers’ expectations to third and fifth grade achievement, and relations of third grade teacher expectations to fifth grade achievement. In addition to the overall effects summarized in Table 13–1, they also found that self-fulfilling prophecies were strongest among minority boys (for reading) and children from low income backgrounds (in math). Clearly, more research is needed to understand the conditions under which students from stigmatized backgrounds are vulnerable to strong self-fulfilling prophecies.

Four studies, however, including Hinnant et al (2009) have addressed —whether positive or negative teacher expectations are more powerful— Unfortunately, the four all provided different, and conflicting results. Hinnant et al (2009) found that self-fulfilling prophecies were entirely linear; that is, positive expectations improved students about as much as negative ones harmed them. This was a high quality study, with a large sample, and rigorous data analysis, so it deserves high credibility. Thus, one possibility is that our results were a sort of random fluke and, upon further replication, it will be found that stronger positive effects are relatively uncommon.

Nonetheless, in addition to focusing on younger children than we did, Hinnant et al (2009) differs in one very important respect from our research. Specifically, it examined effects of teacher expectations across multiple years. Thus, the effects of a particularly inspiring teacher (one who warmly and benevolently expects “too much” from a student) may largely dissipate in a subsequent year when the student has a different teacher. Of course, this is just a speculative possibility, so it will be important for research to examine this issue again, in multiple ways (in the same year, over multiple years) in order to be more confident in the empirically justified conclusions.

Babad et al. (1982) examined the power of negative and positive self-fulfilling prophecies among 26 teachers and 202 students in gym classes who had either low-bias or high-bias teachers (bias referred to degree of cognitive rigidity or dogmatism among teachers). This study reached the conclusion that negative self-fulfilling prophecies were more powerful than positive ones, at least among high-bias teachers. When I looked closely at their results, however, I was not so sure that this conclusion was justified.

Babad et al. (1982) found no differences in athletic accomplishments between high- and low-expectancy students’ athletic performance among low-bias teachers. Thus, there was no evidence of self-fulfilling prophecy at all, neither positive nor negative, among students of low-bias teachers. Therefore, students’ performance among low-bias teachers could be used as a sort of control group for determining whether self-fulfilling prophecies primarily helped or hurt students with high-bias teachers. There were three student performance measures: (1) distance jump, (2) sit-ups (for girls) and push-ups (for boys), and (3) running speed. For the distance jump, negative self-fulfilling prophecies were more powerful than positive ones. Lows with high-bias teachers jumped significantly less far than did lows with no-bias teachers, whereas highs with high-bias teachers jumped the same distance as highs with low-bias teachers. This result, therefore, is consistent with their conclusion that negative self-fulfilling prophecies were more powerful than positive ones.

For sit-ups/push-ups, positive self-fulfilling prophecies were more powerful than negative ones. Low-expectancy students of high-bias teachers performed 3.8 fewer sit-ups/push-ups than did low-expectancy students of no-bias teachers; high-expectancy students of high-bias teachers performed 4.7 more sit-ups/push-ups than did high-expectancy students of low-bias teachers (see their Table 5, p. 469). Although Babad et al. (1982) did not test whether the 4.7 difference was significantly greater than the 3.8 difference, this result disconfirmed the prediction that the effects of negative self-fulfilling prophecies exceed those of positive ones.

The results for their speed measure also provided no evidence of negative expectancy effects exceeding positive ones. The performance of lows with no- and high-bias teachers was similar, indicating that negative self-fulfilling prophecies did not occur. Highs with high-bias teachers actually performed worse than highs with no-bias teachers, which may be an interesting effect of teacher bias but does not represent a self-fulfilling prophecy.

Overall, therefore, Babad et al. (1982) found stronger negative than positive self-fulfilling prophecies for distance jump, stronger positive than negative self-fulfilling prophecies for sit-ups/push-ups, and no self-fulfilling prophecy for running speed. Such results do not seem to justify any general conclusion about the relative power of positive versus negative teacher expectations.

An even earlier study also tested whether positive or negative teacher expectations produced stronger self-fulfilling prophecies (Sutherland & Goldschmid, 1974). Six first- and second grade teachers provided their expectations for each student in their classes 2 months into the school year. Ninety-three students were divided into five teacher expectation groups (ranging from "poor" to "superior"). The students were administered intelligence tests at each of two time points: 2 months and 7 months into the school year.

Sutherland and Goldschmid (1974) first focused on students with below-average IQ scores, who were divided into two groups: (1) those whom teachers believed had average intelligence (erroneously high expectation) and (2) those whom teachers believed had below-average intelligence (accurately low expectation). The self-fulfilling prophecy prediction is that students in the first group (low student IQ/inaccurately high teacher expectation) would show greater increases in IQ over the year than students in the second group (low student IQ/accurately low teacher expectation). The pattern of increases confirmed the prediction for both IQ tests, but the difference was not statistically significant (effect sizes of .1 to .2).

Next, Sutherland and Goldschmid (1974) divided students with above-average IQ test scores into two groups: (1) those whom teachers believed had above-average intelligence (accurately high expectation) and (2) those whom teachers believed had average intelligence (inaccurately low expectations). The self-fulfilling prophecy prediction here was that students in the second group (high IQ/inaccurately low teacher expectations) would show lower increases or greater decreases in IQ test scores than students in the first group (high IQ/accurately high expectations). This prediction was confirmed for both measures; in addition, these differences were both statistically significant and quite strong (r s of .45 to .55).

These results suggest that negative expectations undermined the future IQ scores of high-IQ students, whereas positive expectations had no significant effects on the future IQ scores of low-IQ students. Although these results suggest that negative self-fulfilling prophecies were more powerful than positive ones, this study suffers from several serious methodological weaknesses. First, negative expectations underestimated high-IQ students more than

positive expectations overestimated low-IQ students. Positive expectations consisted of rating as “average” students with IQ scores of 80 to 95. Negative expectations consisted of rating as “average” students with IQ scores of 120 to 135. An average IQ score is 100. Thus, an “average” rating probably underestimates a student with a score of 120 to 135 more than it overestimates a student with a score of 80 to 95.

The greater power of negative versus positive self-fulfilling prophecies that emerged, therefore, may have reflected the greater inaccuracy of negative expectations as operationalized among their particular sample, rather than any generally greater power of negative expectations. More inaccurate expectations have greater potential to be self-fulfilling. Therefore, even if the self-fulfilling effects of teacher expectations in *Sutherland and Goldschmid's (1974) own data* were completely linear (no difference in the power of positive and negative self-fulfilling prophecies), operationalizing teacher inaccuracies in such a manner as to render low expectations more inaccurate than high ones would lead to finding that negative self-fulfilling prophecies exceed positive ones.

In addition, the study did not examine the effects of inaccurately low expectations on low-IQ students or of inaccurately high expectations on high-IQ students. A teacher could believe that some slightly below-average students are even less competent than indicated by their IQ score, or that some high-IQ students are even more competent than indicated by their IQ test score. Such effects, however, were not assessed.

Therefore, this study's results can best be summarized as follows: Highly inaccurate low expectations undermine high-IQ students' future IQ test scores more so than moderately inaccurate high expectations enhance low-IQ students' future test scores. Such a specific and narrow conclusion does not appear to provide a firm empirical foundation for broad conclusions regarding the relative power of positive and negative teacher expectations.

A third study (Alvidrez & Weinstein, 1999) examined the extent to which preschool teacher beliefs about student intelligence predicted the overall high school GPAs of 63 students (all of whom were 4 years old) in the context of a model that controlled for IQ and parental SES, both measured at age 4 (previous analyses showed that neither student gender nor ethnicity predicted GPA beyond the effects of IQ and SES). The results were quite striking: not only did the preschool teacher expectations predict high school grades (overall effect of nearly .4), but also polynomial regression showed that the largest effects occurred for negative expectations (underestimates) and that the effects of positive expectations were near zero.

Why did such a pattern occur? Several limitations to their study render its interpretation ambiguous. First, IQ tests among 4-year-olds lack the reliability and validity of those administered to older people (e.g., Neisser et al., 1996). Furthermore, IQ tests have come a long way since the 1960s, which is when Alvidrez and Weinstein's (1999) data was collected (Neisser et al., 1996).

This raises the possibility that teacher perceptions at age 4 were sufficiently accurate to recognize student characteristics predictive of achievement that were not fully captured by the IQ test. Especially because student grades are often influenced by nonacademic aspects of behavior, such as cooperativeness, disruptiveness, and obedience (Jussim et al., 1998), and because the personality characteristics underlying these behaviors are often strikingly consistent across the lifespan (e.g., Roberts et al., 2007), it is possible that ratings provided by teachers of preschoolers had predictive validity not accounted for by the IQ tests.

Furthermore, Alvidrez and Weinstein (1999) acknowledged many of these issues and clearly stated that their study was not capable of distinguishing between accuracy and self-fulfilling prophecy as explanations for the patterns they observed. We agree, but would go further. They provided no data and little in the way of speculation regarding how the expectations held by preschool teachers for 4-year-old children could actually cause achievement in high school (beyond a general reference to the potential for self-fulfilling prophecies). Far more long-term, longitudinal research is needed before any conclusion that they identified a causal process could be justified (a point they themselves emphasized in their discussion section).

Conclusion

This chapter focused on three questions. How powerful are expectancy effects in the classroom? How accurate is the typical teacher expectation? Have any conditions been identified under which truly powerful self-fulfilling prophecies do occur? Meta-analyses, field experiments, and naturalistic studies all converge on the conclusion that, in general, self-fulfilling prophecies are not very large. As Brophy (1983) suggested over 20 years ago, only about 5% to 10% of students are typically affected by self-fulfilling prophecy effects (most of which fall in the .1 to .2 range). Biasing effects, too, fall in a similarly small range. Self-fulfilling prophecies and biases probably do sometimes combine to create fairly dramatic effects on at least some students' grades. Nonetheless, the only teacher expectation research that has assessed accuracy—the naturalistic studies—consistently shows that accuracy, not self-fulfilling prophecy or judgmental bias, is the main reason students' achievement conforms to teacher expectations.

Although teacher expectancy effects are typically small and accuracy is typically high, the type of large and dramatic self-fulfilling prophecies emphasized by social psychologists do sometimes occur—and in the types of contexts about which social psychologists have expressed the most concern. Specifically, our quest studies showed that genuinely large self-fulfilling prophecies occur among students suffering from some type of stigma—social class stigma, racial stigma, their own histories of low achievement (or, especially, combinations of these stigmas).

Nonetheless, our overall pattern was only partially consistent with the “social problems” orientation of much social psychological research on stereotypes and expectancies. The main inconsistency was that our research indicated that teacher expectations were more likely to be solutions to rather than sources of social problems. Specifically, our quest results showed that positive teacher expectations were more powerful than negative ones—that is, erroneously positive teacher expectations increased student achievement more than erroneously negative teacher expectations harmed student achievement. Determining the generality of this pattern is an important question for future research, especially because of the seemingly different patterns found in other studies (Alvidrez & Weinstein, 1999; Babad et al., 1982; Hinnant, et al, 2009; Sutherland & Goldschmid, 1974—although the degree of difference may be more apparent than real).

It is, of course, possible that some conditions facilitate the occurrence of positive self-fulfilling prophecies and others facilitate the occurrence of negative self-fulfilling prophecies. If so, the sparse evidence on this issue does not yet shed light on just what those conditions

might be. Regardless, it is clear that the issue of whether teacher expectations more strongly undermine or enhance student achievement is an unsettled question. At minimum, however, such an unsettled state of affairs provides no evidence that self-fulfilling prophecies maintain a castelike system by limiting the achievement gains of students from the wrong side of the tracks (see, e.g., Gilbert, 1995; Hofer, 1994; Rist, 1970; Weinstein et al., 2004). Mixed evidence provides no empirical justification for claims emphasizing the power of negative teacher expectations to maintain or exacerbate social problems, inequalities, social stigmas, and the like.

Nonetheless, negative self-fulfilling prophecies are not strictly necessary to maintain the claim that teacher expectations contribute to a castelike system of ever-increasing differences between high- and low-expectancy students. Even if teachers' expectations never harmed students at all, positive expectancy effects, alone, could create ever-increasing differences between high- and low-expectancy students, *if* the same students were the beneficiaries of positive expectancy effects year in and year out. The idea that small expectancy effects may accumulate over long periods of time to become large effects is another conceptual/logical tool in the arsenal of those arguing that expectancy effects are both larger than they seem and a significant contributor to social problems. It is also an important theoretical issue in its own right. Therefore, both theory and evidence regarding whether self-fulfilling prophecies produced by teacher expectations accumulate is examined in the next chapter.

Notes

1. For the statistically disinclined, these are sophisticated statistical techniques that can separate out how much each of two or more predictor variables independently predict such outcome. For example, they might be used to tease out the extent to which teacher expectations predict student future achievement beyond the effects of students' past achievement on their future achievement.

2. Or, more technically, for the statistically inclined, we performed path analyses predicting changes in student achievement (1) using all three of the teacher expectation variables simultaneously but estimating each variable's independent relationship to achievement; (2) using all three of the teacher expectations simultaneously and together (i.e., entering all three together into a regression equation and assessing their combined *r*-squared increment); and (3) using two different LISREL teacher expectation variables: (3a) one assuming a latent teacher expectation variable caused our observed teacher expectation variables and (3b) one assuming that the three teacher expectations reflect (are caused by) an unmeasured teacher expectation variable. Results regarding self-fulfilling prophecies, bias, and accuracy were similar identical no matter how we did it.

3. The statistically inclined may wish to consult Alwin and Hauser (1975) for a discussion of the decomposition of effects in path analysis and Jussim (1991) for a detailed example demonstrating how accuracy mathematically and statistically predictive accuracy equals the correlation minus the path coefficient linking teacher expectations to students' future achievement.

4. These types of explanations are "benevolent" because they blame perceivers, societies, and institutions for creating disadvantage and inequality, rather than the disadvantaged themselves. As such, these types of "benevolent" social explanations stand in sharp contrast to biological explanations—or even explanations based on differing groups' subcultures—which appear to "blame the victim." See Chapter 10 for a more detailed analysis of the role of political ideology in

social scientists' preferences for some types of explanations over others; the role of political ideology in explaining many social scientists' views of stereotypes are also touched on in Chapters 15 through 19 and 21.

5. Actual *Ns* varied considerably from analysis to analysis. *Ns* much below 1,000 are clearly identified in the text of the chapter, though full details are reported in the original research articles (Jussim et al., 1996; Madon et al., 1997).

6. In this chapter, when teacher expectations predicted changes in student achievement, such results are interpreted as self-fulfilling prophecy. This is a causal interpretation ("teacher expectations cause student achievement"). The statistically inclined are probably going bananas about now because everyone who has ever had minimal training in statistics knows that one cannot infer causality from correlation. This chapter nonetheless maintains this interpretation for several reasons. One is for ease and simplicity of writing. If one wishes to read all the scientifically necessary but jargon-laden qualifiers, contortions, and self-flagellations for why causality cannot be inferred from these data, one can read the original articles.

Furthermore, correlations between A and B have three potential explanations: A causes B, B causes A, and C causes A and B (teacher expectations cause student achievement, student achievement causes teacher expectations, or some third factor [or set of factors] causes them both). The longitudinal nature of the data precludes student end-of-year achievement from causing early-year teacher expectations. Student achievement in May of sixth grade cannot possibly cause teacher expectations in the previous September of sixth grade. This leaves two possible sources of correlation between teacher expectations and student achievement: Teacher expectations cause student achievement or something else causes them both. We have, however, assessed and controlled for some of the most likely contenders for "something else causing them both": including prior student achievement and student motivation, and student demographics. Therefore, the argument that the effects are not causal requires one to be suggesting that some other, unassessed factor has caused them both. This is possible, but, as of this writing, no such plausible additional factor has yet been identified.

The fourth reason this causal language is used here can be viewed as a classic case of "be careful what you wish for, you may just get it." If one wishes to argue against a self-fulfilling interpretation of the results reported in the quest studies, one is, in essence, making an accuracy argument. Why? Because if one wishes to suggest that something else causes them both, one is implicitly claiming that teacher expectations have even less self-fulfilling power and are even more accurate than I have (sometimes quite controversially) claimed. Predictive accuracy is predictive validity without causal influence, so this third variable explanation is, in essence, an accuracy explanation. In that case, perhaps there is no evidence anywhere of powerful self-fulfilling prophecies, and even I have understated the case for accuracy! (This would actually strengthen the main themes of this book—that self-fulfilling prophecy and bias are relatively small, and that people are often much more accurate than routinely given credit for being by many social scientists.) So, if you want to take me to task for "inferring causality from correlation," go right ahead—I ask only that you also explicitly state that your interpretation of this research means that "self-fulfilling prophecies are even weaker and accuracy even more powerful than Jussim has concluded."

7. This logic might also be applied to girls, especially in math and science classes. However, we (Jussim et al., 1996) examined whether self-fulfilling prophecies were more powerful among girls than boys and found that they were not.

8. For the statistically inclined, the regression coefficients linking teacher expectations to MEAP scores were .37 and .14 for African American and White students, respectively.

9. For the statistically inclined, “typical” effect here simply means the change in standardized achievement test score (and, by analogy, SAT) associated with a 1-standard-deviation change in teacher expectations. “Largest possible” effect here means the change in standardized achievement test score (and, by analogy, SAT) associated with a 4-standard-deviation difference between having a teacher expectation. A 4-standard-deviation difference was chosen because it reflects the difference between having a teacher expectation that is 2 standard deviations below versus 2 standard deviations above where it should have been based on students’ prior records. Such large discrepancies did not occur often, but they did occur.

10. For the statistically inclined, this section combines and summarizes results from several regression analyses reported in Jussim et al. (1996). This was necessary because two different teacher expectation variables and two different social class variables were involved in the prediction equations.

11. For the statistically inclined, the standardized regression coefficients relating teacher expectations to student achievement among high achievers was .16, whereas among low achievers, it was .24—see Madon et al. (1997) for more details.

12. For the statistically inclined, the standardized regression coefficients relating teacher expectations to student achievement ranged from about -0.10 to about .20 for everyone except low achievers who were targets of high expectations. Among this latter group, however, the standardized regression coefficients were about .4.

13. For the statistically inclined, the standardized regression coefficient relating teacher expectations to the future achievement of lower class, low-achieving students was .62, whereas the same coefficient for other combinations of class and achievement ranged from .13 to .27.

14 Do Self-Fulfilling Prophecies Accumulate or Dissipate?

THIS IS A true story (with some minor details changed) of accumulating negative self-fulfilling prophecies and the hard work and determination that overcame them. Marco (not his real name) grew up on a subsistence farm in northern New Jersey—his family was something about as close to the common cultural image of “Appalachian Whites” as one is likely to find in New Jersey. They were poor. When Marco came home from school, he did not watch TV (they did not even have one till he was in high school). He did not go out to play with his friends. He often did not even do his homework. Instead, he went to work in the fields with his siblings to help his parents try to scratch a meager existence from the small patch of land they farmed.

Young Marco did not do well in school. Eventually, he was diagnosed as having learning disabilities and was placed into special education classes. In practical terms, sometimes such placement is a good thing, even though it reflects in some sense “low expectations.” For children who really do have some sort of bona fide limitation, it is probably better (for maximizing that child’s learning and achievement) to place them with a competent, specially trained teacher than to allow those children to flounder in regular classes (where their disabilities may evoke little more than anger or frustration from a teacher who merely sees them as being lazy, disruptive, etc.).

Nonetheless, it is also possible that such placement is not a good thing. Such placement sometimes means little more than low expectations, low standards, acceptance of shoddy work, and little attempt to push these kids to achieve as highly as possible. In such cases, these classes are probably often a recipe for the types of negative self-fulfilling prophecy effects that contribute to social problems. This was, for the most part, Marco’s experience.

Consequently, Marco fell further and further behind his peers every year. By the time he was in high school, he was, to all outward appearances, what might be colloquially described

as a “loser.” He was from a poor background. He was in special education classes. And he was barely passing even those classes, despite their extraordinarily low standards. He was, frankly, fortunate not to have done anything that could have gotten him locked up.

At this point, no one, including him, considered him to be college material. So, he joined the Navy. One often gets at least two good things out of a stint in the U.S. military: (1) One often learns that through hard work and sufficient time and discipline, one can accomplish many difficult things, and (2) one also receives substantial support from the U.S. government to pay for tuition if one decides to attend college. By the end of his stint in the military, Marco had decided that he did not want to be relegated to the type of life that seemed pre-ordained by his impoverished background and special education public school classes. So, he decided to go to college.

Marco (who, as you probably have guessed, is considerably smarter than all this background would seem to suggest) was smart enough to realize that his high school record precluded him, not merely from places like Harvard or Princeton, but from pretty much any 4-year college. Also, even though he had decided to go to college, the prospect of competing against all those bright and shiny middle- and upper middle class kids—folks with “young inventor” awards and mountains of “extracurricular activities” from high school, who not only had college fully paid for by their parents but also were driving cars that Marco could only wistfully wish for—was a bit daunting. Plus, Marco would have to work his way through college (even though tuition was largely paid for, he still had to eat and pay rent). So, he applied to a local community college, which admitted almost anyone who applied, and began his college career.

Then, something happened. Marco loved college. He loved the world of ideas. And the various strands of his life came together in a stunning and synergistic manner. Marco had always worked hard, whether on a farm or in the military. Now, he applied that same work ethic to his classes. And, out of the blue, Marco—the same Marco who barely passed special education classes in high school—over 2 years worth of classes received nearly straight As in his community college.

Marco, now emerging from his personal academic dark ages, realized that a community college degree was not going to do him a lot of good. Also, although he knew that 4-year colleges were tougher, straight As would be enough to make almost anyone at least a bit more confident in his or her ability to go to the next level. So, he applied to Rutgers. Based on his stellar community college record, he was admitted.

Rutgers is no community college. It is a very tough state school, one in which most students graduated from high school in the top 25% of their class. The average SAT scores are around 1,200 to 1,300—that is, in the top 10% to 20% of all SAT scores. Community college graduates are routinely shocked at how different it is, and it usually takes a semester or two of relatively low grades before they learn how to cope with the difference—if they ever do.

This was only partially true for Marco. He was shocked at how much harder it was. But Marco, by this time, was utterly determined to succeed and knew more than a little about hard work. So, in sports parlance, Marco raised the level of his game and again began pulling nearly straight As in his classes.

At around this point, Marco started working in my lab as an undergraduate research assistant (this is how I have come to know this story). He was quite good, and when he expressed interest in doing an honors thesis in his senior year, I happily agreed. This was no personal

affirmative actionlike decision on my part. Marco was, by this time, simply very good. It was not till much later, when his thesis was mostly done, that he told me all about his background.

And (not surprisingly), Marco did an excellent thesis. His thesis showed that, even when people clearly harbored prejudices against a group, they nonetheless evaluated individuals from that group on the basis of their personal characteristics. Prejudice played no role in their evaluations of those individuals. And Marco went on, not only to graduate from Rutgers with high honors, but also to attend graduate school and complete a master's degree (he is currently a counseling psychologist with a thriving practice).

Marco's story raises all sorts of interesting social psychological issues, especially about the accumulation of self-fulfilling prophecies over time. Although "accumulation of self-fulfilling prophecies" will be described in more analytic detail later in this chapter, for now, let's keep it simple: It means that self-fulfilling prophecies may snowball and have larger and larger effects over time. One interpretation of Marco's story, therefore, might emphasize the power of small self-fulfilling prophecies to accumulate.

When young, he did not do well in school. In a very Rist (1970)-like manner, perhaps part of the reason was the self-fulfillment of unflattering social class stereotypes held by his teachers. Eventually, low expectations for Marco received strong institutional support and approval—he was "labeled" as having learning disabilities. This label, in a very Rosenhan (1973)-like manner, deeply colored most teachers' expectations for Marco—so much so that they all failed to see the potential scholar inside this poorly dressed, poor-performing student. As a result of these consistent low expectations, a downwardly spiraling vicious circle was created, whereby Marco came to dislike and dread school, which further undermined his performance, which teachers could then point to as evidence of his lack of intellectual and academic acumen. So, by the time Marco hit high school, he could barely pass special education classes.

That interpretation is pretty tight. Even if a bit melodramatic or overstated (though no more overstated than the interpretations of similar situations that have appeared in the scholarly literature—see, e.g., Hamilton, Sherman, & Ruvolo, 1990; Rist, 1970; Weinstein, Gregory, & Strambler, 2004), it probably has more than a little truth to it. But it does leave out one important thing. For Marco, as for most of us, life does not end at high school.

One can definitely make a case for the slow but ultimately powerful accumulation of self-fulfilling prophecies for Marco. But one must be careful about just what one concludes, if one makes this case. Because, given that Marco ultimately graduated from a strong college with honors and went on to receive a graduate degree and become a successful professional, it is vividly clear that, whatever self-fulfilling prophecies did occur at one point in his life, their effects had almost completely dissipated a few years later. Thus, another interpretation of Marco's story, one seemingly diametrically opposed to the first one, is that, although self-fulfilling prophecies may occur and may even accumulate over limited periods of time, ultimately, most of the time, they dissipate. According to this interpretation, they dissipate because, in general with many exceptions, you can't keep a good man or woman down. You can act like smart and competent people are stupid and incompetent; you can treat them as if they are stupid and incompetent; you can even make them act like they are stupid and incompetent for a while. But, eventually, most of the time, one way or another, their brains and competence will emerge. Of course, Marco's story is just a story (albeit a completely true

one), and anyone is free to draw whatever message he or she prefers from it. It is not scientific research and so provides no real evidence on which to reach any firm conclusions about whether self-fulfilling prophecies generally accumulate or dissipate. This raises the question, What has the scientific research shown about the accumulation or dissipation of self-fulfilling prophecies? Before discussing the empirical evidence, however, I distinguish between two types of accumulation and review the conceptual arguments suggesting that accumulation is the most likely phenomenon and the arguments suggesting the exact opposite: that dissipation is the most likely phenomenon.

The Logic of Accumulating Self-Fulfilling Prophecies

Many researchers have suggested that empirical studies underestimate self-fulfilling prophecies, because expectancy effects may accumulate over time and/or over multiple perceivers (e.g., Claire & Fiske, 1998; Jones, 1990; Snyder, 1984; Weinstein & McKown, 1998). Furthermore, I have found that the belief that small expectancy effects accumulate over time is very widespread, at least within social psychology—far more widespread than the list of citations at the end of the first sentence of this paragraph might suggest. Routinely, when I present colloquia and research talks emphasizing the limited power of self-fulfilling prophecies, the first questions I get usually go something like this (I have tried to be true to the substance and tone of these questions: Note that they are not really questions, but more like comments that challenge the justification for my conclusion emphasizing small effects):

Even if you are right, the effects may be far larger than you conclude, because targets interact with lots of perceivers, over long periods of time, providing ample opportunity for the effects to become much larger than you report.

Without using the term “accumulation,” this comment implies that the speaker believes that, if we could somehow add together the self-fulfilling effects of lots of perceivers’ expectations, over long periods of time, such effects would be substantially larger than indicated in my presentations that usually point out the relatively weak fragile nature of such effects.

The logic of accumulation, which appears compelling at first glance, is straightforward:

1. Small effects are typically obtained in both short-term (e.g., 1-hour) laboratory studies of self-fulfilling prophecies and teacher expectation studies conducted over a school year.
2. Although small in such contexts, many targets may be subjected to the same or similar erroneous expectations over and over again. For example, students from privileged sociodemographic backgrounds may consistently benefit from high teacher expectations, whereas those from culturally stigmatized backgrounds may be consistently undermined by low teacher expectations. Social stereotypes, assumed to be widely shared and erroneous in many reviews of expectancy effects, are often presented as an obvious reason to predict that targets from stigmatized groups will be subjected to repeated self-fulfilling prophecies from multiple perceivers and over long periods of time (e.g., Claire & Fiske, 1998; Darley & Fazio, 1980; Deaux &

Major, 1987; Fiske & Taylor, 1984, 1991; Jones, 1986, 1990; Jost & Banaji, 1994; Snyder, 1984; Taylor, 1992). Thus, according to this type of analysis, overall effects of expectancies on any particular target are likely to be much higher than demonstrated in any particular study.

The logic of accumulation, when considered by itself, seems compelling—perhaps so compelling as to appear sufficiently obvious or inevitable as to not even require empirical testing. Because it may appear so compelling, this type of conceptual analysis probably contributes to the traditional social psychological emphasis on the power and pervasiveness of self-fulfilling prophecies. But before foreclosing on the need to obtain evidence regarding the conclusion that teacher expectancy effects accumulate over time and across perceivers, it might be worthwhile to consider the social and psychological processes that could work against accumulation.

Potential Limitations to Accumulation

Myriad social and psychological processes might work against accumulation. Although a review of such processes is beyond the scope of this chapter, a few contenders will be briefly mentioned. Within social psychology, perhaps the most obvious is self-verification (reviewed in Chapter 7), which refers to the idea that people are not only often highly motivated to see themselves in a manner consistent with their own long-standing and deep-seated self-views but also often motivated to convince others to view them much as they view themselves. This is important because it means that the self-verification motive may help shield many people from confirming others' inaccurate expectations.

Only one study has addressed both accumulation and self-verification. In a laboratory study conducted over three sessions, Swann and Ely (1984; discussed in detail in Chapter 7) found that, although self-fulfilling prophecies occurred, targets were more likely to convince perceivers to change their expectations than targets were to fulfill perceivers' expectations. Overall, rather than accumulating, the self-fulfilling effects of perceivers' expectations *declined* over the three sessions. Thus, self-verification constitutes one potential obstacle to the relentless fulfillment of others' expectations.

A second potential limitation is accuracy. As people get to know one another, the potential to maintain highly erroneous views of one another may decline (although it probably does not decline to zero; see, e.g., Kenny, 1994). Similarly, popular cultural mythology notwithstanding, rather than rigidly applying stereotypes to every individual who is a member of the stereotyped group, people are typically highly sensitive to individual differences, when those individual differences are experimentally manipulated or readily available in naturally occurring situations such as classrooms (e.g., Jussim, Eccles, & Madon, 1996; Kunda & Thagard, 1996; this issue is discussed in detail in Chapter 18). At least two meta-analyses have shown that stereotype effects on judgments of individuals become progressively smaller the more information perceivers have regarding those individuals (Davison & Burke, 2000; Eagly, Makhijani, Ashmore, & Longo 1991). This means that, even among targets from stereotyped groups, disconfirming behavior is far more likely to be noticed and to influence perceptions and judgments than it is to be ignored or dismissed. Such a process, too, will

typically increase the accuracy of expectations for individuals. If accuracy increases over time, it will limit and reduce the potential for self-fulfilling prophecy.

A third potential limitation is regression to the mean.¹ If a perceiver holds an unusually high or low expectation for a target, even if that expectation is self-fulfilling, the target's behavior may drift back to its pre-self-fulfilling prophecy levels (in the absence of interaction with others who hold equally unusually high or low expectations, which, by definition, will be unusual). Thus, regression to targets' prior levels (of behavior, achievement, etc.) may create a tendency for self-fulfilling prophecies to dissipate, rather than accumulate.

This analysis so far has been conceptual rather than empirical. Such analysis, however, is itself important. The logic of accumulation, by itself, may appear so compelling as to not even require empirical justification. When considered in the context of potentially opposing processes, however, accumulation of self-fulfilling prophecies may not seem to be such an obvious or inevitable outcome of interpersonal interactions. This might bring the issue back from the brink of foreclosure and reopen it as one requiring empirical testing.

Thus, the bottom line is data, not argument. To what extent do the self-fulfilling effects of teacher expectations accumulate? Addressing this question requires understanding two potentially very different types or aspects of accumulation. The first involves accumulation of self-fulfilling prophecies resulting from multiple perceivers within the same time frame (e.g., multiple teachers during the school year). The second involves accumulation of self-fulfilling prophecies over time (e.g., the same teacher over multiple semesters or multiple teachers over multiple years). Each of these is discussed next.

Concurrent Accumulation Effects

The accumulation of self-fulfilling prophecies from multiple perceivers' expectations within a single time period. Within a single time frame (e.g., one school year), the effects on targets of multiple perceivers' expectations may accumulate. To distinguish such effects from the accumulation of expectancy effects over time (e.g., multiple school years), I refer to these as "concurrent accumulation effects" (Jussim et al., 1996). The notion of concurrent accumulation effects is implicit in most perspectives that emphasize the potentially self-fulfilling nature of social stereotypes (e.g., Claire & Fiske, 1998; Deaux & Major, 1987; Hamilton et al., 1990; Jones, 1990; Snyder, 1984). Because stereotypes are often presumed to be both shared and erroneous, perceiver after perceiver will presumably heap self-fulfilling prophecy after self-fulfilling prophecy upon stereotyped targets.

Such a perspective, which is implicit in many discussions of why expectancy effects may be larger than indicated by the empirical evidence, has been explicitly articulated by Claire and Fiske (1998, p. 208):

To understand the significance of the pressure on targets, one must take the perspective of a target across time and interactions. . . . But in constraining possible social influence to short-term one-on-one interactions, the methodology itself [**of studying brief interactions**] reinforces an individualist view of behavior by ignoring the repetitiveness of a target's experience over time and across situations, and the cumulative effect of these interactions.

And later (on p. 211):

Thus, stereotypes are not only widely shared, but some are also pervasively applied in interactions with targets.

The upshot of this analysis is clear: Because all previous research has focused on the potentially self-fulfilling effects of only one perceiver on each target, if multiple perceivers influence targets in daily life, people would be more heavily influenced by self-fulfilling prophecies than is implied by existing research.

Focus on naturalistic, not experimental, studies. My analysis of concurrent accumulation focuses exclusively on naturalistic studies for several reasons. The logic of accumulation across multiple perceivers requires those perceivers to *spontaneously* develop (i.e., not by experimental intervention) similarly inaccurate expectations for a target. If perceivers rarely spontaneously develop similarly inaccurate expectations, there is not even much potential for accumulation in daily life. Contrasting expectations, if self-fulfilling, will negate one another rather than accumulate.

Furthermore, Claire and Fiske's (1998) critique is most fitting for *experimental* studies of self-fulfilling prophecies. Such studies are typically conducted in very narrow contexts—typically a dyadic interaction that takes place over an hour or less. Even the rare exception, such as the long-term field experiment of Rosenthal and Jacobson (1968a, b), could not address concurrent accumulation effects, because false expectations were experimentally manipulated. Assuming the random assignment to condition was successful, there is no reason to think that *other* teachers typically held the same expectations for the “late bloomers” as did those in whom Rosenthal and Jacobson (1968a, b) instilled false expectations. Exactly as Claire and Fiske (1998) argued, therefore, such studies do ignore “the repetitiveness of the target's experience over time.”

In contrast, the typical naturalistic study often provides an appropriate context for studying the accumulation of concurrent self-fulfilling prophecies. Most naturalistic studies of teacher expectations, for example, are conducted over at least one school year, thereby allowing for the possibility that multiple teachers will develop similar expectations for students, and at least raising the possibility of concurrent expectancy effects.

A faulty implication. Unfortunately, however, perspectives suggesting that concurrent accumulation effects are likely to be larger than suggested by the existing literature are seriously flawed. They draw a faulty implication from two sound premises. The two sound premises are:

1. Existing research on self-fulfilling prophecies has focused on interactions between two people (teacher–student, employer–employee, etc.) of limited duration.
2. To the extent that targets interact with many perceivers who share expectations, the cumulative effect of self-fulfilling prophecies will exceed the effect occurring in any given interaction.

This leads to the seemingly self-evident but nonetheless false conclusion that:

3. Studies focusing on interactions of limited duration underestimate the extent to which targets are affected by self-fulfilling prophecies, because such studies cannot measure the self-fulfilling effects of all perceivers *not* included in the study.

How can point 3 be false, when points 1 and 2 are true? The answer is that point 3 describes the actual state of affairs precisely backward: Studies focusing on dyadic interactions *do not underestimate* expectancy effects from multiple perceivers; instead, such studies *overestimate* effects of individual perceivers' expectations precisely because they (unintentionally) incorporate the effects of all other perceivers with overlapping, self-fulfilling expectations! To understand how this could be requires a brief statistical detour.

A brief but necessary detour: Omitted variables in naturalistic studies. This is a self-fulfilling prophecy variant on the well-known "omitted variable problem" in regression. For the statistically uninitiated, regression is a statistical technique in which one or more variables are used to simultaneously predict some outcome. For example, one might use height and weight to predict strength, or SATs and high school GPA to predict college GPA. The outcome of a regression indicates how much each hypothesized predictor predicts the outcome, after controlling for all other predictors in the model (is strength based more on height [controlling for weight] or more on weight [controlling for height]? To what extent do SATs [controlling for GPA] and high school GPA [controlling for SATs] predict college GPA?).

The omitted variable problem in regression refers to the fact that if one has missed (i.e., not included in one's analyses) some third variable (or variables) that causes or is strongly associated with both the predictors and the outcome included in one's regression, the results will be inaccurate. Specifically, if one has omitted such an important variable, the results will probably indicate that the included variables more strongly influence the outcome than they really do. For example, if relentlessly working out at the gym influences both weight and strength (if workout frequency is omitted from predicting strength), the influence of weight on strength may be overestimated. If social class influences both SAT scores and college GPA (if social class is omitted from predicting college GPA), the influence of the SATs may be overestimated.

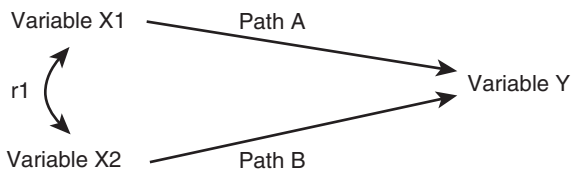
If this brief delving into statistical arcania has not lost you or put you to sleep, perhaps you are getting an inkling of where I am going with this. In short, the expectations held by anyone not included in one's study are "omitted variables"—variables which may inflate the estimate of expectancy effects in naturalistic studies. Precisely how is discussed next.

Other perceivers' expectations as omitted variables: Why existing studies overstate self-fulfilling prophecy effects of single perceivers' expectations and precisely estimate multiple perceiver effects. Nearly all existing studies of self-fulfilling prophecies focus on dyadic (two-person) interactions. Even in classroom studies, where there might be dozens of teachers and hundreds of students, the analyses focus on determining the effect of a particular teacher's expectations on a particular student's achievement. Although such effects may be reported for the whole sample, the level of analysis is the dyad—the two-person or teacher–student unit.

Now, here is something that I found to be amazing when I first realized it was true. Even though all existing studies of naturally occurring self-fulfilling prophecies focus exclusively on dyadic interactions, *they implicitly assess the self-fulfilling effects of all perceivers holding expectations similar to those of the perceiver included in that study. All perceivers. Even the expectations of the potentially great number of perceivers not included in the study.* The idea that a study can assess effects of perceivers' expectations not included in that study, at first glance, might appear inconceivable and perhaps even nonsensical. It is true nonetheless and Figure 14–1 shows why.

Figure 14–1 shows a simple model presenting a hypothetical example in which two different teachers hold expectations for a set of students. Path's A and B depict the causal,

Model 1: General Model



Model 2: Two Teachers Hold Similar Expectations

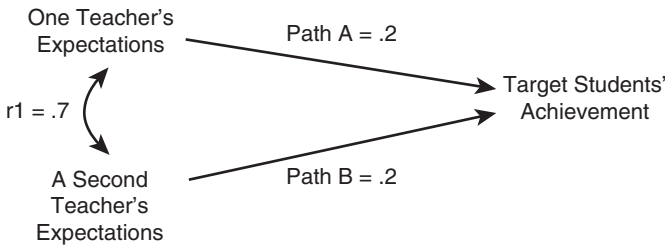


FIGURE 14-1 Concurrent Self-Fulfilling Prophecy Accumulation Effects

self-fulfilling effects of each teacher's expectations. The self-fulfilling effects for each teacher's expectations are both .2. $r1$ is the correlation between the two teachers' expectations. It equals .7, which means that the two teachers hold very similar, but not identical, expectations for their students.

This analysis shows ways in which those speculating that expectancy effects are more powerful than they seem are both correct and incorrect. If two perceivers hold similar self-fulfilling expectations, there will be more of an expectancy effect than if there is only one perceiver holding self-fulfilling expectations. In this sense, those arguing for accumulation are correct. In this concrete hypothetical example, the total self-fulfilling effect of the two teachers' expectations is .34, not .2. The statistically curious can see endnote 2 for a full mathematical explanation.²

But the conclusion that empirical research underestimates expectancy effects because it fails to account for multiple perceivers is incorrect. The Figure 14-1 model demonstrates that such a conclusion is topsy-turvy: Studies focusing on dyadic interactions do not *underestimate* cumulative expectancy effects from multiple perceivers. Instead, such studies *overestimate* effects of individual perceivers' expectations because they (unintentionally) incorporate the cumulative effects of all other perceivers with self-fulfilling expectations similar to those of the perceivers in the study. Excluded perceivers are *omitted variables* whose omissions artificially inflate the expectancy effect obtained for the included perceiver.

In the example displayed in Figure 14-1, if one *only* assessed the self-fulfilling effects of Teacher One, one would obtain an effect of .34. This overestimates Teacher One's self-fulfilling prophecy effect. It does not underestimate the total effect of Teachers One and Two, *even though Teacher Two's expectations were not assessed*.

Omitted variables positively correlated with two measured variables will almost always artificially *increase* the size of the assessed relation between the included variables. Thus, the

effects of expectations similar to those of the perceiver in the study, as held by the potentially almost infinite number of perceivers excluded from any particular study, are omitted variables. These omitted expectation variables may inflate the size of the assessed relationship of the perceiver's expectations to the target's behavior or achievement. In other words, concurrent accumulation effects are already (implicitly) assessed in dyadic studies of naturally occurring social interactions, such as between teachers and students.

The empirical evidence on the power of concurrent accumulation effects. I know of only one study (described later) to have directly empirically assessed the cumulative self-fulfilling effects of multiple perceivers' expectations. Nonetheless, the models in Figure 14-1 clearly show that we already have clear evidence about the general extent of such effects. Concurrent accumulation effects exactly equal the extent and power of self-fulfilling prophecies as assessed in naturalistic studies of dyadic interactions. Such studies overestimate dyadic self-fulfilling prophecies precisely to the extent that concurrent accumulation occurs. In some sense, this may be a "flaw" in those studies, in that they probably overestimate dyadic self-fulfilling prophecies. But with respect to understanding the likely extent of concurrent accumulation effects, this "flaw" is a boon—it means that all we have to do is examine existing naturalistic research to identify the likely extent of concurrent accumulation effects.

This analysis means that the research reviewed earlier showing that teacher expectation effects are typically around .1 to .2 probably overestimates the effects of individual teachers' expectations. But such results are probably a reasonably good estimate of the *total* effect of all teachers' (i.e., including those not in the study) expectations that are similar to those of the teachers actually included in the study. Although this review of conceptual issues and empirical evidence cannot conclusively demonstrate the existence of concurrent accumulation effects, it does conclusively demonstrate that, *if* they occur, they are fully captured by the .1 to .2 self-fulfilling prophecy effect sizes typically found in naturalistic studies of teacher expectations.

Synergistic accumulation of parents' beliefs about their children's alcohol use. Madon, Guyll, Spoth, and Willard (2004) proposed that parents' erroneous expectations could have synergistic self-fulfilling effects on their children. Specifically, they suggested that when parents held highly similar but erroneous expectations for their child, those expectations would synergistically enhance each other's effects, to produce a particularly powerful self-fulfilling prophecy. This is a type of concurrent accumulation, because it involves accumulation across two perceivers (parents). However, it differs from the type of concurrent accumulation discussed so far, because it involves two people's expectations not merely adding self-fulfilling prophecies on top of one another, but actually enhancing one another's power (which is why Madon et al. called this "synergistic accumulation").

Madon et al. (2004) examined this issue with respect to parents' beliefs about their seventh graders' likelihood of drinking alcohol. Overall, the self-fulfilling effects of parents' beliefs were typically small (.05 for dads and .16 for moms). However, these overall figures popped up to .26 for moms and .15 for dads *if* both mom and dad similarly overestimated their child's likely drinking. Although even these effects, individually, are not all that powerful, remember that these parents' expectations were similar, so that both produced similar self-fulfilling prophecies. And, when added together, their effects were substantial—at least twice as many of these kids reported drinking alcohol as did the other children.

Interestingly, however, this pattern only occurred when parents erroneously overestimated their child's alcohol use. Underestimates produced no similar synergistic effect. Nonetheless, these results at least raise the possibility that, when targets are faced with a uniform phalanx of others' erroneous expectations, especially if they are powerful others with whom the target has a long-term relationship, self-fulfilling prophecies may be considerably larger than usual.

Concurrent accumulation: Conclusions. In general, concurrent accumulation effects in the classroom are not very powerful. Still, the Madon et al. (2004) article at least raises the possibility that, sometimes, erroneous expectations may synergistically create self-fulfilling prophecies that are more powerful than usual. Of course, that is a single study, focusing on a single context. Whether synergistic accumulation actually occurs, and how frequently, in classrooms, boardrooms, assembly lines, ball fields, and therapists' couches will remain an unanswerable question pending research in these contexts. Nonetheless, these patterns naturally lead to the next question: Do self-fulfilling prophecies accumulate over time?

Accumulation Over Time

The accumulation-over-time hypothesis is that a self-fulfilling prophecy process triggered by a perceiver's expectations at one time continues so that targets conform more and more to the perceiver's original expectations over time. A perceiver's initial false belief more strongly influences targets over time. Thus, the impact of self-fulfilling prophecies may transcend the original context of the interaction and profoundly influence targets (Claire & Fiske, 1998; Snyder, 1984).

The logic of accumulation over time is, at first glance, as compelling as that of concurrent accumulation:

1. Self-fulfilling prophecies clearly occur in limited contexts, such as the lab or school year.
2. Small effects may accumulate over multiple school years so that initially small differences between high- and low-expectancy students become large.

For example, consider two students both starting sixth grade with IQs of 100. Suppose that the sixth grade teacher believes that one of these students is bright and the other is not. Also assume that teachers' expectations have an effect of .2 on student achievement (an effect of .2 is equivalent to one fifth of a standard deviation and the standard deviation of IQ tests is 15). Thus, by the end of sixth grade, the "bright" student's IQ will be 103 and the "dull" student's IQ will be 97. If this small effect accumulates over time, then by the end of high school, the "bright" student will have an IQ of 115 and the "dull" student will have an IQ of 85. In this example, each year from 6th through 12th grade, the gap between the low- and high-expectancy students widens by 3 points. Thus, small expectancy effects have the potential to become much more powerful via accumulation.

Again, however, such an analysis is compelling only in the absence of a comparable analysis of factors likely to limit accumulation over time. Self-verification, accuracy, and regression all may limit accumulation over time. Furthermore, different teachers in different school years may not hold equally inaccurate expectations for most students.

For example, consider a B student who is the target of an erroneous high expectation in a ninth grade math class. Through self-fulfilling prophecies, such a student receives a final grade of B+ in that class. To accumulate, the tenth grade teacher would have to have *another erroneously high expectation*, which is also self-fulfilling, such that the student would end up with an A. If the tenth grade teacher expected a B+, regardless of whether one considers this accurate or self-fulfilling, and the student receives a B+, there is no accumulation. There is merely a self-fulfilling prophecy effect that occurred in ninth grade that was sustained, not increased, in tenth grade.

Again, however, the bottom-line issue is the data, not conceptual analysis. What has research shown regarding the accumulation of self-fulfilling prophecies in the classroom? To date, four studies have addressed this issue, and they are discussed next.

Dissipation in Pygmalion

The classic Pygmalion study (Rosenthal & Jacobson, 1968a, b) was also the first to address the accumulation issue. Even if one uncritically accepts their results, they clearly showed that self-fulfilling prophecies did not accumulate. Bloomer-control differences at the end of Year 1 were about 4 IQ points; such differences were less than 3 IQ points at the end of Year 2. Thus, self-fulfilling prophecies appear to have started dissipating (although whether they would have dissipated to zero is a question that cannot be answered by their data, because they only followed students for 2 years). Rosenthal and Jacobson (1968a, b) did not, however, test whether the bloomer-control IQ differences at the end of Year 1 were statistically significantly larger than those at the end of Year 2. Thus, all that can be claimed is that their pattern supported dissipation, not that they had statistically significant evidence of dissipation. Clearly, however, their evidence disconfirmed the accumulation hypothesis.

The Permeability of Social Class in the Classroom

The Rist (1970) observational study described earlier in this book (see Chapters 4 and 6) also addressed the accumulation issue because he followed the class of students from kindergarten to second grade. Because Rist's study did not include quantitative measures of student achievement, he used table assignment as a criterion for identifying the self-fulfilling effects of teachers' expectations on students. Even if one accepts table assignment as a criterion for identifying self-fulfilling prophecies, his results provided some evidence of dissipation and no evidence of accumulation (and if one does not accept table assignment, one must conclude the study provided no evidence that bears on the accumulation issue). At each transition to a new grade, some students were moved between supposed competence levels (see Chapter 6). This is evidence of both stability and dissipation (stability for those who stayed in the same group; dissipation for those who moved up or down). Thus, despite Rist's (1970) emphasis on how teachers created a social class-based "caste" system, the actual results, based on table assignment, provided evidence of both stability and dissipation of self-fulfilling prophecies from kindergarten through second grade. Of course, this was an observational

study, and no statistical comparisons regarding degree of stability or change were reported. Again, however, the pattern suggested that dissipation was occurring to at least some extent, and there was no evidence of accumulation.

Dissipation in Early Elementary School ---

The Hinnant et al. (2009) study described in Chapter 13 also found evidence for dissipation. About 600 to 750 students (depending on the analysis) were followed from first through fifth grade. Whereas first grade teacher expectations had typically small effects on third grade reading achievement (0 to about .2), they had no significant effects on reading whatsoever by fifth grade. Similarly, whereas first grade teacher expectations had a substantial effect on third grade math achievement (.36), this effect was reduced to .09 by fifth grade. Hinnant et al. (2009), however, did not test whether the decline in these effects were themselves statistically significant. Thus, the extent to which this supports dissipation is unclear. Clearly, however, accumulation did not occur.

Nonetheless, that effects of first grade teacher expectations may manifest up to four years later is itself interesting and important. Although such effects are quite modest, and might even eventually dissipate to zero (as did the reading effects), the longterm duration of self-fulfilling prophecies, even in small, diluted form, testifies to the important impact they can have on students.

Dissipation Through High School ---

West and Anderson (1976) analyzed data from 3,000 male students in their freshman, sophomore, and senior years of high school that included information on both teachers' expectations and student achievement. The effect of teachers' expectations on sophomore-year achievement was .12, whereas the effect on senior-year achievement was .06. These results appear to support the dissipation hypothesis. Teachers' expectations from freshman year predicted senior-year achievement less strongly than they predicted sophomore-year achievement. West and Anderson (1976), however, did not assess whether .12 was significantly larger than .06. Thus, the extent to which this supports dissipation is unclear. Clearly, however, accumulation did not occur.

Incomplete Dissipation From Sixth Grade Through High School ---

We also directly examined this issue as part of our "quest" for the powerful self-fulfilling prophecy. Specifically, we (Smith, Jussim, & Eccles, 1999) examined whether teacher expectation effects accumulated from 6th through 12th grades (*Ns* ranged from about 500 to 1,700, depending on the analysis). Outcomes included both final grades and standardized test scores. The main results showed no evidence of accumulation. For the most part, self-fulfilling prophecies dissipated. Teacher perceptions in sixth and seventh grade predicted both grades and standardized test scores more weakly over time.

Although the results predominately supported the dissipation hypothesis, we also found that the expectancy effects in 1 year were very long-lasting. That is, teacher perceptions in sixth and seventh grade predicted significant changes in student achievement through high school. In the sixth grade analyses, the strength of the relationship between teacher perceptions and student standardized test scores declined from about .1 in sixth grade to near zero by twelfth grade; the relationship with final marks declined from about .35 in sixth grade to .17 in twelfth grade. In the seventh grade analyses, the strength of the relationship between teacher perceptions and student standardized test scores dropped off from .16 in seventh grade to .09 in 12th, and the relationship with final marks dropped off from almost .5 to about .25. All of these declines were statistically significant.

Although self-fulfilling effects did not accumulate, they did have long-lasting consequences. The evidence of self-fulfilling prophecies occurring in sixth and seventh grade continued to manifest, albeit in considerably diluted form, through all of high school. Thus, like Hinnant et al.'s (2009) results, the durability of such effects—over many years and many different teachers—was quite striking, even if the overall power of such effects was quite modest.

Attractiveness and the Salary of People with MBAs

Frieze, Olson, and Russell (1991) examined accumulation outside the classroom—they examined it among people with MBAs on the job. Frieze et al. (1991) summarized their study as showing that “underlying beliefs about people based on their physical appearance affect their judgments and behaviors toward those individuals” (p. 1053). My view is that this study provided evidence of accumulation that could best be described as tantalizing—they showed something accumulated, but whether it was self-fulfilling prophecy was unclear. Specifically, Frieze et al. (1991) examined whether (1) attractive people with MBAs would receive higher starting salaries than unattractive people with MBAs and (2) these differences increased over time. Starting salaries were predicted from attractiveness, years of full-time work prior to receiving the MBA, and the year in which they took their first full-time job after receiving the MBA (for the statistically inclined, this was done using regression).

Their results. Did attractive people receive higher starting salaries? Sort of. The classic social psychological analysis would seem to predict that, if anything, attractiveness would be a greater predictor for women than for men, given the nature of traditional sex stereotypes and the male-dominated world of business. But it wasn't so. Attractive men received significantly higher starting salaries than unattractive men and attractiveness made no difference among women. Frieze et al. (1991) speculated that this might be because attractive men and unattractive women are seen as more masculine than, respectively, unattractive men and attractive women, and masculinity is seen as desirable for managerial jobs.

Additional analyses showed that physical attractiveness significantly predicted subsequent salaries for both men and women. The average salary difference between attractive and unattractive men more than doubled over time, from a starting difference of about \$2,200/year to a difference of about \$5,200/year after several years on the job. Much of this difference, however, was not due to self-fulfilling prophecy or any other social process occurring on the job. How is this knowable? In another set of analyses, Frieze et al. (1991) controlled for initial

starting salary. This is important because, given two different starting salaries, identical percentage increases will increase the difference in ending salary.³ Attractive men did increase their salary differential over unattractive men, but this primarily resulted from their initially greater starting salaries. There was no significant difference in the later salaries of attractive and unattractive men, after controlling for the differences in their initial salaries.

Even though there was no initial starting salary difference between attractive and unattractive women, several years later, the attractive women were making about \$4,300/year more. Frieze et al. (1991) interpreted this as a self-fulfilling prophecy, and they could be correct. However, several aspects of this study undermine the confidence we can have in this conclusion.

Did they really find self-fulfilling prophecies? It is not clear. First, they never surveyed the employers. Thus, there was no assessment of employers' expectations for individual employees. Therefore, whether salary differences were caused by employers' physical attractiveness stereotypes is unknowable. Without evidence regarding employers' expectations, it is somewhat difficult to conclude that those expectations created a self-fulfilling prophecy. Perhaps they were caused by something else.

Are there any likely contenders? There are, because the study likely suffers from at least one omitted variable problem. There was no measure of the skills, competencies, or performance at other jobs of the people with MBAs. How might this be relevant? More attractive adults are in fact more socially skilled than less attractive adults (Eagly, Makhijani, Ashmore, & Longo, 1991; Feingold, 1992). Because social skills might be crucially important in many managerial positions, more socially skilled people with MBAs (who are, on average, more attractive) might deserve and receive higher salaries than less socially skilled people with MBAs (who are, on average, less attractive).

Of course, perhaps attractive young adults became socially skilled through self-fulfilling prophecies occurring before they took their first job. As discussed extensively in Chapter 10, however, if they were more socially skilled when they applied for the job, then they deserved to be seen as more socially skilled. If they were already more socially skilled *at that time*, and if social skill was relevant to the job, then their receipt of higher pay could have resulted from an accurate assessment of their higher level of competence rather than a self-fulfilling effect of employers' attractiveness stereotypes.

Other limitations to the study further undermine the viability of the self-fulfilling prophecy interpretation. For an attractiveness-based self-fulfilling prophecy to have occurred, the job performance of the attractive people with MBAs should have improved more so than did the job performance of the unattractive people with MBAs. But Frieze et al. (1991) did not assess the job performance of the people with MBAs. This leaves open myriad possibilities, in addition to accuracy. Perhaps the employers' expectations biased their judgments of the people with MBAs without changing their objective performance, and these judgments led to increased salaries (much like teacher expectations may bias their judgments of students and lead to increased grades even in the absence of a self-fulfilling prophecy). Perhaps the attractive people with MBAs were more likely to sleep their way to the top and received higher salaries.

Clearly something happened—because the more attractive people with MBAs had higher salaries. Although I have criticized their study for not really providing evidence to back up their preferred self-fulfilling prophecy interpretation, it should be noted that I have cited no

evidence in their study to support the accuracy interpretation, the bias interpretation, or the sleeping-to-the-top interpretation. That's because the study provided no evidence capable of distinguishing among these explanations.

Well, that's not completely true. They actually did provide some hints of accuracy. Specifically, the strongest predictor of starting salary was years of work experience prior to receiving the MBA, and the strongest predictor of final salary was years of full-time work since receiving the MBA. In other words, employers heavily rewarded experience, which seems awfully close to evidence of accuracy. Nonetheless, I think it is fair to characterize their results as tantalizing.⁴

Conclusions Regarding Accumulation

There might be some conditions under which teacher expectation effects accumulate to create large, enduring differences between students. Such conditions, however, have not yet been found. Despite arguments and claims emphasizing the power of expectancy effects to accumulate, there is no evidence of powerful accumulation effects in the classroom.

Concurrent accumulation most likely occurs, at least sometimes. The path models in Figure 14-1 show that whenever two or more teachers hold similar expectations for students (or, more generally, perceivers for targets), if those expectations produce self-fulfilling prophecies, such effects will accumulate. However, despite the initially compelling appeal of arguments for the concurrent accumulation hypothesis, naturalistic studies of dyadic teacher-student interactions, which typically show expectancy effects of .1 to .2 (see Chapters 3 and 13), do not underestimate concurrent accumulation. Instead, they precisely capture the overall extent of accumulation across different teachers during the time period covered by those studies. Concurrent accumulation effects in the classroom are, therefore, on average quite small.

Interpreting the evidence regarding accumulation over time is more straightforward. At least in the classroom, it does not happen. Five studies have directly addressed this issue (Hinnant et al., 2009; Rosenthal & Jacobson, 1968a, b; Rist, 1970; Smith et al., 1999; West & Anderson, 1976). None found evidence of accumulation. All found at least some evidence that self-fulfilling prophecies dissipate. The four that followed students for more than 2 years, however, also found that, although self-fulfilling prophecies dissipated, they generally did not evaporate completely (Hinnant et al., 2009; Rist, 1970; Smith et al., 1999; West & Anderson, 1976). Thus, although there is no evidence of accumulation effects, there is good evidence that self-fulfilling prophecies that occur in 1 year can have long-lasting consequences.

Outside the classroom, the evidence on accumulation is still preliminary and suggestive. Madon et al.'s (2004) results suggested that parents' expectations for their children's alcohol use combine in a synergistic manner; Frieze et al.'s (1991) results suggested that the physical attractiveness stereotype might be self-fulfilling at work. Self-fulfilling prophecies probably do, at least sometimes for some people under some conditions, accumulate. And such events are both theoretically interesting and practically important. The current evidence, however, strongly suggests that such effects are unusual, rather than common or powerful (e.g., most combinations of parents' expectations did *not* accumulate, even in the Madon et al. [2004] study).

Furthermore, despite the fact that one can tell a really compelling story about how the accumulation of self-fulfilling prophecy upon self-fulfilling prophecy constitutes a major mechanism by which social stereotypes confirm themselves and maintain unjustified systems of oppression and status (e.g., Claire & Fiske, 1998; Darley & Fazio, 1980; Snyder, 1984; Weinstein et al., 2004), there is, in fact, currently no clear evidence supporting such an analysis, and a great deal of evidence disconfirming it (see all the teacher expectation studies demonstrating dissipation).

This brings us back to stereotypes, a topic touched on in many prior chapters but not yet fully addressed in much depth. That, however, is the purpose of the next several chapters.

Notes

1. For the statistically disinclined, regression to the mean sounds very technical, but the idea is quite simple. It refers to the fact that, after some extreme occurrence, things typically (not always) return to their ambient prior average levels. For example, consider a Tuesday in February in which it is 60 degrees in Anchorage Alaska. Although it is possible that it will be 65 or 70 degrees on Wednesday, it is far more likely that the temperature will fall below 60 degrees on Thursday—that is, fall back in the general direction of the mean temperature in Anchorage in February. Similarly, consider a student who just received a grade of B+ on a midterm, despite consistently having received Cs in most prior schoolwork. Such a student is more likely to receive a B or a C+ on the next exam than to receive an A. This is probability, and, of course, nothing is guaranteed; exceptions occur. Regression, however, need not occur if some systematic process has been put in place that will create a change in the overall mean. If global warming becomes sufficiently severe, it may eventually be common for 60 degree days in February in Anchorage to be followed by 65 degree days. If our hypothetical student has all of a sudden seen the academic light, has dramatically altered his or her study habits, and now works vastly harder to achieve excellence, maybe an A is more likely than a C+. In the absence of some systematic push in a particular direction, however, the best prediction is that, following some extreme occurrence, things will be most likely to return to their prior normal levels.

2. Understanding this idea requires a very basic understanding of decomposition of effects in path analysis (Alwin & Hauser, 1975). For simplicity and to illustrate the basic ideas, I have used a very simple three-variable model. The principles, however, apply identically in more complex contexts involving more complex models.

In Figure 14-1, the total correlation between One Teacher's Expectations and Target Students' Achievement comes from two sources. The first is the causal effect of One Teacher's Expectations (Path A = .2). If one variable causes another, they will correlate with one another. Thus, the .2 causal influence of One Teacher's Expectations on Target Students' Achievement causes .2 worth of correlation between One Teacher's Expectations and Target Students' Achievement.

But there is another, noncausal, source of correlation between One Teacher's Expectations and Target Students' Achievement that derives from their mutual associations with A Second Teacher's Expectations. One Teacher's Expectations are correlated with A Second Teacher's Expectations ($r_1 = .7$); Target Students' Achievement is also associated with A Second Teacher's Expectations (Path B = .2). In this simple model, figuring out how much this increases the correlation between One Teacher's Expectations and Target Students' Achievement is quite easy—just multiply

$r_1 \times \text{Path B}$. $r_1 \times \text{Path B} = .14$. So, the total, overall, zero order correlation between One Teacher's Expectations and Target Students' Achievement is .14 more than it would be if Path A was the only source of correlation. Therefore, the overall correlation between One Teacher's Expectations and Target Students' Achievement equals: $\text{Path A} \times (r_1 \times \text{Path B}) = .2 + .7 \times .2 = .34$.

If one failed to include A Second Teacher's Expectations in the model, Path A becomes the zero order correlation between One Teacher's Expectations and Target Students' Achievement, and Path A would equal .34. Thus, unincluded perceivers' expectations are merely *omitted variables*. The failure to include them leads to upwardly biased estimates of dyadic self-fulfilling prophecy effects. But the omitted variable "problem" in regression constitutes a wonderful gift to researchers interested in the accumulation of self-fulfilling prophecies, because it means that, even if studies overestimate dyadic effects, they fully capture all concurrent accumulation effects!

3. Consider the following hypothetical example. Attractive Arnold starts his job earning \$50,000/year and Unattractive Unger starts at \$40,000/year. So, their starting salary difference is \$10,000. They each receive a 5% raise for 5 consecutive years. At the end of 5 years, Attractive Arnold is earning \$63,814 and Unattractive Unger is earning \$51,051. So now the difference is almost \$13,000. Thus, even in the complete absence of self-fulfilling prophecy, the salary difference increased over time, purely as a result of the difference in initial starting salary. This is why it was crucial for Frieze et al. to control for initial starting salary.

4. For the statistically inclined, Frieze et al. (1991) reported only unstandardized regression coefficients and *t* values; no information was provided regarding means, standard deviations, correlations, or standardized coefficients. This renders it impossible to compare their results to those of other studies or to compare the extent of self-fulfilling prophecy versus accuracy.

This page intentionally left blank

6 Stereotypes

This page intentionally left blank

15 On the Pervasiveness and Logical Incoherence of Defining Stereotypes as Inaccurate

BEFORE CONTINUING WITH the rest of this chapter, please take the following test.

1. In the United States, members of which group are most likely to commit murder?

Men	Women
-----	-------

2. In which ethnic/racial group in the United States are you likely to find the highest proportion of people who supported Democratic presidential candidates in 2000 and 2004?

Whites	African Americans
--------	-------------------

3. People in the United States strongly identifying themselves as _____ are most likely to attend church on Sunday.

Conservative	Liberal
--------------	---------

4. Rank order the following U.S. racial/ethnic groups according to their annual household income.
(first = most):

African Americans	Asians	Whites
-------------------	--------	--------

5. An _____ newspaper claimed that the Indian Ocean tsunami that killed over 150,000 people in 2004/2005 might have been caused by _____ nuclear testing.

A. Israeli; Pakistani	B. Egyptian; Israel
-----------------------	---------------------

6. This is a true story. On December 24, 2004, a dad and his three kids wandered around New York City around 7 p.m., looking for a restaurant, but found most places closed or closing. At the same time, his wife (their mom) performed a slew of chores around the house. This family is most likely:

Catholic	Baptist	Jewish	Pagan/Animist
----------	---------	--------	---------------

7. Please match up the country on the left with the common type of self-concept found in that country on the right.

Great Britain	Collectivist (close friends and family are viewed as part of the self)
Japan	Individualist (the self is unique and independent of others)

The correct answers appear in endnote 1.¹

If you got at least one question right, perhaps you do not need to read this chapter. This is because (1) all these questions assess either the accuracy of your belief about a group (questions 1, 2, 3, 4, 7) or your ability to use your knowledge about groups to make an accurate prediction about an individual or small group of individuals from that group (questions 5 and 6); (2) if you got at least one right, you now have your own personal *prima facie* evidence that all beliefs about groups—all stereotypes—are not necessarily wrong, irrational, and malevolent; and (3) this chapter presents a spirited intellectual critique of many of the reasons stereotypes are routinely viewed as wrong, irrational, and malevolent. This critique will help justify a definition of stereotypes that makes no assumptions about their (in)accuracy and, therefore, constitutes one of the foundations for considering the question of the accuracy of stereotypes as one requiring empirical investigation.

(A colleague suggested that “one right” is too low a standard because people will get three or four right, on average, purely by chance. He, however, did not read Chapters 10 through 12. *Why* your beliefs may be accurate is completely irrelevant to demonstrating their accuracy. If you guessed that the dad in New York City story involved Jews, you were right. You correctly identified that family’s religion. The end. A single exception disproves any absolute, as in, “A single accurate stereotype means that stereotypes cannot be defined as inaccurate.” But I digress. . . . If you are reading this chapter without reading the rest of the book, you might consider the full “digression”—Chapters 10 through 12 vigorously contest much of what I consider to be the confused, unjustified, and often disparaging claims about accuracy so frequently found in the social psychological literature.)

What This Chapter Does and Does Not Address

This chapter focuses exclusively on defining “stereotype.” It will take a whole chapter to do this because so much myth, misconception, Sturm, and Drang have surrounded stereotypes and stereotyping.

To understand how to define stereotypes, therefore, I believed it was necessary to first wade through the swamp of knee-jerk assumptions, politically benevolent but scientifically incoherent claims, and outright preachiness that characterize so much of what so many people seem to believe about stereotypes. So, before presenting my definition, I will address moral, logical, and conceptual issues involved in understanding the (in)accuracy of stereotypes.

The central issue addressed in this chapter is whether stereotypes are inaccurate by definition. Many laypeople, scientists, and scholars alike seem to think they are. I don’t. This chapter, therefore, does not review the empirical evidence regarding the (in)accuracy of stereotypes. That is left for the next several chapters. Instead, it reviews the history of how the conceptualization and understanding of what a stereotype is has changed over the last 90 years; it presents, discusses, and rejects a slew of definitions that have been proposed over the years; and it provides, explains, and justifies a very simple definition that does not assume anything about the accuracy or inaccuracy of stereotypes.

Organization of This Chapter

This chapter has three sections, which I describe here in reverse order. The very last section presents, explains, and justifies my definition of stereotype. This definition does not assume they are accurate or inaccurate; it leaves the issue of (in)accuracy as an open question requiring research, rather than “answering” it by definition.

Such a definition, however, is unlikely to make sense to anyone steeped in the “stereotypes are inaccurate, unjustified, irrational, etc.” political and intellectual tradition. It can only begin to make sense after addressing some of the well-known but largely unjustified beliefs about stereotypes, which is the job of the second section of this chapter. The idea that stereotypes are inaccurate or “unjustified” pervades both social science scholarship and the wider culture for some very good reasons. There are numerous historical examples of stereotypes being used for horrible propaganda purposes to justify and perpetuate some of the most awful cases of oppression, such as slavery, ethnic cleansing, and genocide. Unfortunately,

however, despite the political benevolence of those who condemn stereotypes as irrational tools of oppressors, there are major logical and conceptual problems with taking seriously any assumption that stereotypes must be inaccurate. Given the prevalence of the stereotype inaccuracy assumption, the second section of this chapter presents a critical analysis of this assumption and highlights many of the logical and conceptual problems that emerge when one assumes that stereotypes are inaccurate by definition.

The idea, however, that it is unreasonable to define stereotypes as inaccurate is not merely alien to some people; it is out and out anathema. That some stereotypes might sometimes be reasonably accurate clearly threatens many people. Thus, even before discussing the logical incoherence involved in defining stereotypes as inaccurate, this chapter has a first, preliminary section that tackles these sorts of emotional, irrational, and defensive reactions head-on. This first section has the following purposes: (1) to briefly document the intense hostility the idea that merely raising the possibility of stereotype accuracy evokes in some people, (2) to briefly document the pervasive extent to which the idea of stereotype inaccuracy has been promoted by some of the most prestigious and influential social psychologists, and (3) to thoughtfully consider the morality of allowing versus not allowing for the possibility that some stereotypes may be accurate.

This first section accomplishes this by juxtaposing, first, very real situations in which very real people (including some social psychologists) have, essentially, in lay parlance, “freaked out” at a serious consideration of stereotype accuracy against, second, the wide variety of situations in which either laypeople, experts, or even social psychologists themselves take for granted the reasonableness, appropriateness, and morality of believing that groups differ (i.e., holding a justified stereotype). Of course, if it is even *sometimes* reasonable and moral to believe that groups differ, it cannot possibly be immoral to consider the possibility that some stereotypes might have some degree of accuracy. This chapter also considers the morality of assuming stereotypes are false without testing for their accuracy, and it considers the morality of purposely turning a blind eye to bona fide group differences. This first section, then, discusses and vigorously contests the view that there is something immoral about considering the possibility that some aspects of some people’s stereotypes may be accurate.

Part I: Is It Immoral to Even Suggest That Some Aspects of Some Stereotypes May Be Accurate?

MY UPPER MIDDLE CLASS JEWISH IN-LAWS EXPLODE

Some years ago, my wife’s parents (who are Jewish) asked me what research I was working on. I said, “I was studying the accuracy of stereotypes.” They said, “That sounds like it could be interesting, but could you be more concrete?” “Sure,” I said. “For example, Jews really are, on average, richer than other people.” ****BOOM****. They exploded. They indignantly attacked me for being the worst kind of bigot. They accused me of being anti-Semitic. They accused me of perpetrating the worst type of propaganda about Jews.

This went on for almost an hour. In addition to it being a difficult and tense conversation, however, it did have its ironic side. At the time, their income put them in the top 5% to 10% of all Americans, and their net worth was about \$2.5 million, also placing them well within

the top 5% of Americans. Obviously, their relative affluence did not necessarily mean Jews in general were affluent. But, in the back of my mind (the front being occupied with making peace), this delicious irony was readily apparent.

Eventually, I figured out what was going on. When I said, “Jews really are, on average, richer than other people,” that is not what they heard. What they heard was “Jews are all a bunch of cheap, corrosive, money-grubbing vermin who should be exterminated.” Once I understood that that was what they heard, their reaction made sense, and I was more able to defuse their intense hostility.

I asked them if they were proud of the fact that the proportion of Jews with advanced and professional degrees and in prestigious professions, such as law, medicine, architecture (my father-in-law is an architect), and science, greatly exceeded the proportion of Jews in the population. My asking a leading question that portrayed Jews in a favorable light also seemed to calm them down a bit. They, being proud of their Jewish heritage, readily agreed. They specifically traced Jewish success in America to the long-standing Jewish cultural emphasis on education. “Good point,” I said, and then added, “do you think all those Jews with advanced degrees make the same money as high school dropouts?” They said they did not know, but I urged them to think about it. I then asked them if they believed that all those Jewish doctors, lawyers, architects, and scientists generally earned income comparable to that of, say, janitors, bus drivers, cashiers, cab drivers, and waiters.

Slowly, painfully, I could see a grudging dawn of recognition come into both their faces and into this discussion. The narrow claim that Jews were wealthier than other folks, although it could be used for nasty propaganda purposes, was, itself, a statement of fact, not a racist libel. If one looked at it from a standpoint of respect and accolade for Jewish culture, they much more easily accepted it. Putting it simply and bluntly, though—“Jews are richer than other people”—sounded to them like a harsh indictment of Jews.

MY SOCIAL PSYCHOLOGIST “IN-LAWS”

I have told this story because I think it may help convey why many people have a viscerally hostile reaction to scholarship that seriously considers the possibility that some stereotypes may be accurate. And this seems to be just as true of many social psychologists as it was of my in-laws, many of whom seem to harbor more than a little hostility to the idea of stereotype accuracy.

Sixty years of empirical research has told us much about stereotypes. Stereotypes can arise from, and sustain, intergroup hostility. They are sometimes linked to prejudices based on race, religion, gender, sexual orientation, nationality, and just about any other social category. They can serve to maintain and justify hegemonic and exploitative hierarchies of power and status. They can corrupt interpersonal relations, warp public policy, and play a role in the worst social abuses, such as mass murder and genocide. For all these reasons, many social scientists—and especially many social psychologists—have understandably approached stereotypes as a kind of social toxin.

Perhaps equally understandable, but scientifically untenable, is the corresponding belief that because stereotypes sometimes contribute to these many malignant outcomes, they must also be—in the main—inaccurate. The tacit equation is, If stereotypes are associated with social wrongs, they must be factually wrong. However, the accuracy of stereotypes is an

empirical question, not an ideological one. And for those of us who care deeply about stereotypes, prejudice, and social harmony, getting to the truth of these collective cognitions should guide inquiry about them.

Unfortunately, this has not always been my experience. Because of my research into stereotype accuracy, I have been accused by prominent social psychologists of purveying “nonsense,” of living “in a world where stereotypes are all accurate and no one ever relies on them anyway,” of calling for research with titles like “Are Jews really cheap?” and “Are Blacks really lazy?,” of disagreeing with civil rights laws, and of providing intellectual cover for bigots.² When I am on professional research panels and participate in symposia, the only time I have ever received overtly hostile comments or questions has been when I have given talks on stereotype accuracy (not, e.g., when I give talks on self-fulfilling prophecies, the biases produced by stereotypes and prejudice, etc.).

These reactions are understandable, if one remembers that social psychology has a long intellectual history of emphasizing the role of error and bias in social perception, and that nowhere has this emphasis been stronger than in the area of stereotypes. Just consider the following quotes ranging across several decades, including many by some of the most influential social psychologists of their times:

“However, a great deal of the thrust of stereotyping research has been to demonstrate that these behavioral expectancies are overgeneralized and inaccurate predictors of actual behavior of the target individual” (Darley & Fazio, 1980, p. 870).

“The term stereotype refers to those interpersonal beliefs and expectancies that are both widely shared and generally invalid” (Miller & Turnbull, 1986, p. 233).

“The large literature on prejudice and stereotypes provided abundant evidence that people often see what they expect to see: they select evidence that confirms their stereotypes and ignore anomalies” (Jones, 1986, p. 42).

“The problem is that stereotypes about groups of people are overgeneralizations and are either inaccurate or do not apply to the individual group member in question. . . categorization can lead to oversimplification and distortion. . . . In such instances, people tend to perceive members of the other group as all alike or to expect them to be all alike, which they never are” (American Psychological Association, 1991, p. 1064, emphasis in original).

“In this section of the paper, we consider some representative findings to illustrate the powerful effect of social stereotypes on how we process, store, and use social information about group members” (Devine, 1995, p. 476).

“Research has shown many ways in which stereotypes, like a dangerous virus, can survive and perpetuate themselves despite attempts to eradicate them. They can bias the interpretation of a target person’s behavior and generate assumptions about that person in the absence of any real evidence, all in line with stereotypic content. . . . Moreover, they can do so automatically, behind the perceiver’s back so to speak, so that he or she will have no chance to correct the situation. . . . [W]e do not believe that conscious control over the effects of activated stereotypes are that likely to occur outside of the laboratory . . .” (Chen & Bargh, 1997, p. 557).

“Assigning identical characteristics to any person in a group, regardless of the actual variation among members of that group” (Aronson’s, 1999, definition of stereotype, p. 307).

“... stereotypes are maladaptive forms of categories because their content does not correspond to what is going on in the environment” (Bargh & Chartrand, 1999, p. 467).

“... overgeneralized sets of beliefs about members of a particular social group” (Schultz & Oskamp’s, 2002, definition of stereotype, p. 63).

“A stereotype is any generalization about a group. . . . By definition, a generalization about a group is bound to be ‘unjustified’ for some portion of the group members” (Nelson, 2002, p. 5).

“Expectancies exist in the eyes of beholders and actors. As such, disconfirmation of expectancy resulting from stigma and stereotyping is very difficult” (Niemann & Maruyama, 2005, p. 415).

“Even when there is a ‘kernel’ of truth to a stereotype, stereotypes are typically stronger and more pervasive than the kernel would justify (S. T. Fiske, 1998), presumably because the strength and consistency of a phenomenon are exaggerated in perceivers’ minds, augmented by processes such as selective attention, selective exposure, and selective recall” (Hall, Coats, & LeBeau, 2005, p. 914).

To enter this zeitgeist and to argue for the need to take seriously the possibility that sometimes, some aspects of some stereotypes may have some degree of accuracy, therefore, is to risk making claims that are unbearable to some social scientists. But science is about validity, not “bearability.” It is about logic and evidence. This chapter presents the logic part. Chapters 16 through 18 review the evidence.

There are some sharp contrasts in the conclusions reached here as compared to what can usually be found in the typical social psychological discussion of the evils of stereotyping. This chapter and the next two, therefore, may indeed be a source of discomfort—and most certainly a source of disagreement—among anyone who has bought the “stereotypes are inaccurate, irrational, and rigidly resistant to change” view that has dominated discourse on stereotypes for almost a century. Furthermore, over the last 30 years or so, many social psychologists have come to reap large fees by testifying in court on behalf of plaintiffs bringing antidiscrimination lawsuits. And there seems to be no countervailing force—I am not aware of a single social psychologist who has ever testified as an expert witness for a defendant accused of discrimination (they probably exist, but they are few and far between compared to those testifying for plaintiffs). Therefore, many social psychologists now have a vested economic interest in promoting a view of stereotypes as inaccurate, unjustified, major culprits in discrimination, and pervasively harmful.

Thus, political goals, a scholarly tradition emphasizing error and bias (see Chapter 10), and the large fees available in antidiscrimination lawsuits all push social psychologists to emphasize the nasty and error-prone effects of stereotypes. Of course, someone is not wrong about something just because he or she has a vested scientific or economic interest in it. Ideally, the primary source of scientific conclusions about some phenomenon would be scientific data. Thus, disagreements about the accuracy of stereotypes can, presumably, be resolved by a careful approach to defining our terms (to make sure we all mean the same thing when we use a term such as “stereotype”) and by data. It is in this spirit of clearly defining terms and sticking close to the data and letting the chips fall where they may that this and the next three chapters have been written.

STEPPIN' FETCHIT, BOJANGLES, THE HAPPY HOUSEWIFE,
AND THE GRASPING, HOOK-NOSED JEW

Figure 15–1 presents some classic media “stereotypes” over the last century or so. In movies prior to about 1950, African Americans were routinely depicted as ignorant, gullible, superstitious, and musical. The “Happy Housewife,” a media depiction of women as pathologically obsessed with cooking and cleaning, and who are completely fulfilled by such activities, should be familiar to anyone who has ever seen a television commercial or magazine ad from about 1950 to about 1980 (they still appear, but perhaps not quite so frequently). And the dark, hook-nosed, sinister, grasping Jew should be familiar to anyone who has ever been exposed to anti-Semitic propaganda. Many other groups have also been depicted in similarly offensive ways (relentless depictions of Arabs as terrorists in movies, South American fascists and drug dealers, and so on).

1930s media portrayals of African Americans



Steppin' Fetchit



Bojangles

The 1950s “Happy Housewife”



The dark, grasping, hook-nosed Jew



FIGURE 15–1 Classic Media Portrayals of “Stereotypes”

Such images are indeed offensive and the moral outrage they sometimes evoke is well-justified. They also fit the classic view of stereotypes as extreme, exaggerated, unjustified, etc. But what do they actually tell us about stereotypes? Well, the answer to that question depends on what a stereotype is. In general, social psychologists' primary goal is to understand how people think and feel about, and interact with, other people. We study laypeople, "normal" people (at least in the sense that we usually do not study mental illness and we do not restrict our studies to aristocrats or elites). So, a stereotype has to be some sort of belief held by everyday walking-around normal people.

Unfortunately, these media images, no matter how nasty and no matter what their impact, do not tell us anything directly about laypeople's beliefs. Stepin' Fetchit's roles tell us something about movie directors and writers from the 1930s, but exactly what is actually unclear. Do they tell us what their stereotypes were? Or do they merely tell us what they thought would make for a movie likely to sell well? The Happy Housewife? Same deal—she tells us something about what the ad men (and they were overwhelmingly men) of the 1950s and 1960s thought would help sell product. The hook-nosed Jew? Something about how demagogues and propagandists pursue their nasty ends.

These are all important, but they do not tell us anything directly about what the cab driver, the college student, the lawyer, the real estate broker, or the teacher think about various groups. I conclude, therefore, that even though the study of media depictions of social groups is worthwhile on its merits, it actually tells us very little about what everyday people think about social groups. To find out what everyday people think about groups, we cannot study movie directors, advertising executives, or propagandists. We have to study everyday people. For this purpose, such media images are not very useful.

THE STORY OF ROXBURY PREPARATORY SCHOOL

Roxbury Prep is a Boston charter school.³ It's located in one of the poorest, most troubled, most disadvantaged areas of Boston—Roxbury—which is an African American and Latino community racked by crime, drugs, and poverty (it also serves two other similarly impoverished sections of Boston—Dorchester and Mattapan). The students attending the regular public schools in Roxbury perform as poorly as do students in similar districts around the country—high absenteeism, low standardized test scores, high dropout rates, and rare matriculation in college.

The founders of Roxbury Prep wanted to change that. So they created a school for families who aspired to something more. They decided to focus on middle school, because they wanted their children to have a bona fide chance to go to good colleges, and they believed that if they waited till high school, it would be too late. So they created a school for sixth to eighth graders.

One hundred percent of the students attending Roxbury Prep are minorities: about 80% African American and about 20% Latino. About two-thirds participate in the federal Free and Reduced Lunch Program, which provides decent meals in school for children from impoverished families. About two-thirds of the students also enter the school performing a year or more below grade level in reading and math.

Amazingly, however, this school succeeds at elevating the achievement of its students—to levels little short of astonishing. On the Massachusetts Comprehensive Assessment System

(a state-wide series of standardized tests on English, math, and science administered to sixth through eighth graders), in 2004–2007, Roxbury Prep students performed at levels *higher* than those of the average White student in Massachusetts. Indeed, their performance was at a level comparable to some of the most affluent districts in the state.

How did they do it? Although I doubt they would have described it this way, it is clear that they started with stereotype accuracy (beliefs about the ethnic, cultural, and personal characteristics of the groups of children living in their district). Founders and administrators developed a curriculum informed by understanding the unique needs, backgrounds, and experiences of the cultural and demographic groups they hoped to serve. They recognized that these groups of students were seriously behind. They further recognized that, to catch up, they were going to have to work twice as hard.

They created a demanding environment of challenge and high standards (rather than one of remediation). They required far more work and time from their students, which included the following:

- 20% longer school day
- Double periods for math and English (and if there was ever a literal operationalization of “working twice as hard to catch up,” this is it!)
- Readings and songs celebrating African and African American history and culture (not instead of, but in addition to, the “western canon,” which they could pull off because of the double periods)
- *Mandatory* after-school enrichment classes

Why did they design a school that required so much *more* work from its students? *Stereotype accuracy!* The school’s founders recognized (understood, correctly believed, etc.) that the students living in Roxbury, Dorchester, and Mattapan who would attend the school were, in general, likely to be seriously behind other students around the state. The school was several years in the making, so, of course, the founders did not know in advance precisely which students would attend. Therefore, they designed a school to serve the needs of the general profile of the type of student living in Roxbury/Mattapan/Dorchester (yes, this is a sort of benevolent “profiling”). Furthermore, to design a school solely around the needs of the first entering class of students, *if those students were completely unique and different from all other students*, would have been extremely silly, because the school would then have become obsolete for the next entering class. Roxbury Prep’s extraordinary emphasis on working twice as hard makes sense only because Roxbury Prep’s founders designed the school to meet the needs of the types of students *typically or generally* living in its area. In other words, they started with stereotype accuracy!

In some ways, it may be easier to see how this is stereotype accuracy by contrasting what they did with what they *did not* do.

1. They did not ignore the fact that this group of students had different needs and backgrounds than middle class White kids. There was no knee-jerk egalitarianism that denies real differences.
2. They did not dismiss as racists people who claimed that this group of students was underperforming.

3. They did not claim that one would be committing an immoral act of bigotry by presuming that future cohorts of entering students would also be behind and underperforming.

If one wishes to start ameliorating social inequalities, as many social scientists and other people of good will do, one *must* start with a willingness to acknowledge, recognize, and admit that there *is* a particular inequality. In my view, failing to do so is far more immoral than is doing so.⁴ Of course, I know of no social scientists who deny inequalities, which is my main point here: If *some* beliefs about groups are justified—including some that, out of context, might seem highly disparaging (e.g., “those kids are failing”)—then it cannot possibly be scientifically appropriate to define *all* such beliefs as inaccurate.

UNDOCUMENTED SPECULATIONS ON THE NEGATIVE EFFECTS OF DOCUMENTING ACCURACY IN STEREOTYPES

Nonetheless, explicit attempts to deny or dismiss stereotype accuracy have periodically appeared in the scientific literature. There are two broad classes of arguments against taking the possibility of stereotype accuracy seriously. One class is purely scientific and these claims boil down to the idea that it is not methodologically possible to do so (e.g., Fiske, 1998, 2004; Stangor, 1995). Readers are directed to Chapters 10 through 12 of this book, where they will find these methodological arguments energetically contested (see also Judd & Park, 1993; Ryan, 2002). A second class is essentially political—that research documenting accuracy of stereotypes could be used by bigots to promote their evil agendas (Stangor, 1995; see also Fiske’s [1998] claim that such research implies disagreement with civil rights law). This political objection, too, was addressed at length in Chapter 10 (see also Jussim, 2005).

It has yet to happen. Here I make one additional point. This suggestion currently remains a purely hypothetical fiction that seems to primarily reflect the fears of those opposing research on stereotype accuracy. Moving away from hypothetical nightmares and returning to facts, there has yet to be a single documented situation in which the literature demonstrating some accuracy in stereotypes has been used by malicious people for evil purposes.

On the irrefutability of hypotheticals. Of course, I cannot deny the possibility that research on stereotype accuracy could be misused by people to exacerbate social problems. Nor can you deny the possibility that research on stereotype accuracy will be used by people you view as good and decent for socially benevolent purposes. That is the thing about hypothetical possibilities—good or bad, most are irrefutable.

If we reject anything that “might” be used for evil, what is left? It might be helpful to keep in mind, however, that nearly all research can be used for good or evil. Consider research on evolution (social darwinism), rocket science (the Nazis’ V1 missiles lobbed at London), and attitude change (advertising exploiting people’s weaknesses and fears to sell unhealthy products), and just about any other research. If the argument that “this research could be used for evil purposes” is a valid one for not conducting research in an area—in the utter absence of evidence documenting such evil—then the upshot would be that all scientific research should be halted, because it all can be potentially misused. This argument, if taken seriously and applied writ large (and not just to stereotype accuracy research), boils down to a recommendation that we return to the stone age, although even that is probably not far enough

(our hominid ancestors undoubtedly used rocks as tools, including, perhaps, for throwing at and hitting one another). Bad argument, in my opinion.

“But, isn’t the risk higher?” “Wait,” you may be thinking, “the issue is not one of absolutes; it is not that we *know* that demonstrating accuracy in stereotypes will help fascists, sexists, and racists. Seriously, though, you have to admit or at least consider the possibility that there is a greater *risk* of such research doing social damage than, say, research on other, more anti-septic topics.” Well, I have considered it and, although I am open to changing my mind, for now, reject it. One reason not to consider the risk particularly great is that, if it was, there should have been more realization of that potential. That has not happened.

A second issue is that potential costs do not occur in a vacuum. So, in figuring out how to proceed we cannot consider only the potential risks or costs associated with taking stereotype accuracy seriously. In addition, we have to consider the costs and benefits of taking stereotype accuracy seriously versus the costs and benefits of unrealistically assuming there are no group differences when there obviously are such differences. Such lists are too long to review here, so I bring up only three. First, psychology’s “problem” focus is so pervasive that a mini-movement—called positive psychology—has emerged to try to counteract the unduly stark, dark view of human nature produced by psychology. If people are not good at something, we do want to know. But if they are good at something—including perceiving group differences—and we refuse to acknowledge it, we psychologists are contributing to creating an unjustifiably negative and systematically distorted psychology. And scientists are not permitted the luxury of maintaining demonstrably false beliefs once the data come in.

Second, many attempts to deny stereotype accuracy seem to reflect, in part, a genuine concern with redressing social inequalities. Doing so, however, requires first acknowledging the existence of those inequalities. Recognizing, for example, that groups differ in their academic achievement, cultural practices, political beliefs, and economic status is one manifestation of stereotype accuracy. Those arguing that stereotype accuracy cannot or should not be assessed implicitly assume that their beliefs about inequality are accurate (a stereotype!) and, worse, implicitly, have erected a barrier to the very social progress they hope to promote. Acknowledging inequality means recognizing one type of group difference, but, if we are morally prohibited from acknowledging group differences, how can we address that inequality?

Becoming what they fear. This quickly gets to the third and, for me, core cost of denying stereotype accuracy without evidence. Any science, if it wishes to be credible, cannot be in the business of denying reality, even if it is for some supposedly greater political purpose. If it does get into this business, it has shed its role as an objective explorer of the nature of reality and has, instead, become a propaganda tool serving a particular political agenda. Those attempting to squash or deny stereotype accuracy for political purposes, therefore, are acting in a manner much like those they fear in their nightmares—censoring research because it does not fit their politics (if reviewers or editors refuse to publish work on stereotype accuracy because they consider it “offensive,” they are, functionally, censoring such work). For me, that cost is far too high.

Furthermore, in contrast to the purely hypothetical social damage that taking stereotype accuracy seriously can do, the classic social psychological emphasis on the evils of

stereotyping has created real, not imagined, damage. A documented case of this is discussed in the next section.

DOCUMENTED HARMFUL EFFECTS OF EMPHASIZING STEREOTYPE BIAS WITHOUT CONSIDERING STEREOTYPE ACCURACY: A CASE STUDY

Bill von Hippel is a social psychologist. We went to grad school together, and even though we now live half a world apart (he is in Australia), I consider him a friend. He is a good and decent person, is a superb tennis player, and has done interesting and important research on stereotypes and other topics. Like most social psychologists, he strives to use our field's research to reduce social ills. And, like most social psychologists, he presents in his classes a view of stereotypes that I would describe as the "classic" view: as an evil reflecting and causing prejudice, discrimination, injustice, and inequality.

At least, that is how one of his young, earnest, hungry-for-knowledge-and-insight undergraduates once interpreted his lectures on stereotypes and prejudice. And, one night, she needed some cash. So, she hesitated only for a moment as she approached an ATM with several young, male, African American hoodlum-looking types hanging around it. Remembering Bill's lectures, and the fundamentally erroneous, evil, and prejudicial nature of stereotypes, and determined to be the kind of good, decent, egalitarian person whose life is not tainted by bigotry, she walked up to the ATM and proceeded to withdraw her cash—at which point she was mugged and robbed.

As Bill (von Hippel, 2004) tells this story, there is an epilogue. When the student told Bill about this event, their conversation went something like this (this is a re-creation to convey the gist, not the actual conversation):

BILL: If they dressed and acted the same, but were White, would you have withdrawn the cash?

STUDENT (hesitating only a moment): No.

BILL: Well, then, the problem was that you reacted to them on the basis of their race, not on the basis of the individuating information, available from their demeanor and appearance cues, and, of course, it is almost always better to judge others on the basis of their individuating information, rather than their social categories.

This is a great point that will be explored in Chapter 18 in some detail. For now, though, my point is only this. Even if Bill's reply and analysis are 100% correct, it means that the victimized student became a victim because she misunderstood his lecture. In other words, misunderstanding the well-intentioned, egalitarian-motivated, "we are all bigots and have to constantly fight our own bigotry" view promoted by social psychology has now been documented to cause at least one student to unnecessarily become a victim of crime. This experience is familiar to me: Frequently, when I tell people that stereotypes are a mix of accurate and inaccurate (meaning that they are far more accurate than usually given credit for being), variations of this "a person was trying to be unprejudiced and got mugged" story start coming out of the woodwork.

Bottom line: There is now published evidence that those promoting the "classic" view of stereotypes (as irrational, prejudicial, etc.) have caused at least one instance of social harm,⁵

and there remains no published evidence that work on stereotype accuracy has caused social harm.

So, if the criterion for deciding whether an area of research should be abandoned is its tendency to cause social harm, there is now more basis for abandoning research on the error-producing and biasing nature of stereotypes than there is for stereotype accuracy. I am not actually recommending this; I believe most of that research is worthwhile and has done plenty of social good. I am, however, pointing out another example of hypocrisy among those making this case. “Potential for causing harm” appears to be a good argument for abandoning stereotype accuracy research to some researchers. The same argument, however, can be applied at least as readily, if not more so, to stereotype inaccuracy research. Indeed, this is probably merely a subset of the general point that ignorance, distortions, and half truths cause more damage than do knowledge and truth.

Understanding when stereotypes are inaccurate, irrational, and causes and reflections of bigotry versus accurate, rational, and reasonable guesstimates under uncertainty is terribly important if we wish to identify when stereotypes cause harm, when they reflect a reasonable view of reality, and which common beliefs need to be changed. Doing so, however, is impossible, if one is not allowed to study the accuracy of stereotypes to find out which are accurate and which are inaccurate.

More Hypocrisies in the Denial of Stereotype Accuracy

ON THE NEED TO REFUTE A STRAW ARGUMENT

No social scientist has ever explicitly claimed anything quite as silly as “all beliefs about groups are inaccurate.” Thus, demonstrating the silliness of such an argument might appear to be refuting a straw argument. And perhaps it is. But if it is a straw argument, there is an awful lot of straw lying around:

1. For decades, stereotypes were predominantly defined as inaccurate, with virtually no evidence demonstrating inaccuracy (see reviews by Brigham, 1971; Mackie, 1973; Ryan, 2002).
2. There are few, if any, statements in the scientific literature identifying “THESE” types of beliefs about groups as stereotypes, but “THESE OTHER” beliefs about groups as not stereotypes. The hundreds of studies investigating the role of target group memberships in people’s memory, judgment, attribution, perception, and evaluation have routinely been framed as studying “stereotypes” (see reviews by Ashmore & Del Boca, 1981; Brigham, 1971; Dovidio & Gaertner, 2010; Fiske, 1998; Fiske & Neuberg, 1990; Kunda & Thagard, 1996). That research has addressed people’s beliefs about racial, ethnic, religious, social class, gender, national, occupational, and college groups (e.g., dorm residences or colleges attended) or individuals from those groups. Research framed as studying stereotypes has addressed beliefs about political groups (such as Democrats and Republicans), college majors, and “day people” and “night people” (see Chapters 16 through 18 for examples of each of these). Although some researchers have restricted “stereotypes” to beliefs about personality traits (e.g., Brigham, 1971), most do not (see reviews by Ashmore & Del

Boca, 1981; Ryan, 2002). Research framed as addressing “stereotypes” has addressed beliefs and perceptions about behavior, personality, attitudes, criminal culpability, and competencies. As far as I can tell, operationally, the social sciences (as a field; I am not referring to individuals here) have considered people’s beliefs about any attribute regarding any type of group to be a stereotype. It seems, then, that, for all practical purposes, the social sciences consider any and all claims and beliefs about groups to be stereotypes.

3. Putting points 1 and 2 (from the prior paragraph) together:

Point 1: Stereotypes are inaccurate.

Point 2: All beliefs about groups are stereotypes.

Therefore, the inexorable logical conclusion of this line of academic reasoning, whether explicitly stated or not, is that all beliefs about groups are inaccurate.

Again, no researcher has ever made such an absurd claim. Instead, what happens is far more subtle:

4. Among those who define stereotypes as inaccurate, statements of what sort of beliefs about groups are accurate (and, therefore, not stereotypes) almost never appear (for concrete examples, see, e.g., Aronson, 1999; Bargh & Chartrand, 1999; Campbell, 1967; Devine, 1995; Jones, 1986, 1990; Miller & Turnbull, 1986; Schultz & Oskamp, 2000). This, then, opens the door for researchers to consider any and all beliefs about groups to be stereotypes (Allport, 1954/1979, remains a lone exception).
5. Even when researchers do not *define* stereotypes as inaccurate, inaccuracy is often reimported via the “back door”—as something bad, immoral, and unjustified that should be stamped out or avoided (discussed at length later in this chapter).

It is, of course, impossible to know what all these researchers “really believe.” All that is knowable is their published scholarship. And, collectively, the traditional social scientific emphasis on the inaccuracy of stereotypes, combined with its history of considering any belief about any group to be a stereotype, combined with its collective failure to clearly delineate what *is not* a stereotype, appears to inexorably lead to the conclusion that all beliefs about groups are inaccurate.

Perhaps the inexorable logical conclusion that social psychology, collectively, implicitly assumes that all beliefs about groups are inaccurate is itself made of straw. At minimum, this discussion highlights the need for researchers who define stereotypes as inaccurate (overgeneralized, rigid, etc.) to clearly state the criteria for deciding when a belief about a group is *not* a stereotype. Absent that, they leave themselves open to the interpretation that they assume that all beliefs about groups are stereotypes and that, therefore, they assume that all beliefs about groups are inaccurate. Since Allport (1954/1979), however, researchers have almost never declared what sort of beliefs about groups are *not* stereotypes. In this context, therefore, I think that refuting the argument that all beliefs about groups are inaccurate will provide an important contribution to the social sciences, if only by motivating those who define stereotypes as inaccurate to clearly delineate when a belief about a group is accurate and therefore *not* a stereotype.

LOGICAL INCOHERENCE

An explicit claim that all beliefs about groups are inaccurate is, of course, worthy of ridicule on purely logical grounds. It would mean that:

1. Believing that two groups differ is inaccurate
and
2. Believing two groups do not differ is inaccurate.

Both 1 and 2 are not simultaneously possible, and logical coherence is a minimum condition for considering a belief to be scientific. On logical grounds alone, therefore, we can reject the (straw?) argument that all beliefs about groups are inaccurate. Such logical incoherence is usually a red flag that something other than pure science has gone into the assumption of stereotype inaccuracy.

OF MICE AND STEREOTYPES

When a mouse is a research subject, scientists are bound by a set of rules and regulations requiring them to treat the mouse in as moral and humane a manner as possible. They need to be fed regularly (except when studying the effects of hunger *per se*). They need to be kept in clean cages. They can be sacrificed for good scientific purposes, but they cannot be sacrificed gratuitously.

Under very slightly different conditions, however, the mouse does not have the same rights. When a snake is a research subject, it, too, must be cared for in as moral and humane a manner as possible, which, of course, includes feeding it regularly. Snakes eat mice. When a mouse is merely food, the same ethical rules do not apply to that mouse as to the research subject mouse (Herzog, 1988).

What does this have to do with stereotypes? More than it seems. With both mice and stereotypes, change the context, and the morality changes.

In contexts in which beliefs about groups constitute “stereotypes,” such beliefs are frequently assumed to be a nearly unmitigated evil. Such situations include but are not restricted to graduate and undergraduate classes in the humanities and social sciences on stereotypes, prejudice, discrimination, intergroup relations, and social problems; research papers, chapters, and books on the same topics; casual discussions of the same topics; public and political discussions of inequality; almost any time a public figure refers to group differences as having a biological basis and, sometimes, merely when they refer to group differences; and almost any time one person believes another’s beliefs about groups are unjustified.

There are, however, many nonstereotype contexts in which group differences are so obviously real that people take for granted the reasonableness of believing in group differences. If there are any contexts in which it is reasonable to take group differences seriously, then stereotypes cannot always be inaccurate. These contexts are discussed next.

THE ACCEPTANCE OF GROUP DIFFERENCES I:
EXPERIMENTS AND SPORTS TEAMS

In both of the following situations, people (including many of the same social scientists who treat any belief about groups as stereotypes and decry the inaccuracy or unjustified nature of

stereotypes) accept as a matter of course the reasonableness, appropriateness, and morality of believing that groups differ:

1. Social scientists conduct experiments on people and reach conclusions based on average differences between experimental and control groups.
2. Sports fans conclude that championship teams are better than teams that do not perform well enough to even play in championships.

“Wait,” you say, “those comparisons are not fair. First, people do not hold stereotypes of such groups. Second, in those situations people are not born into the groups, plus there is clear and objective evidence that one group really is better than or different than another.”

I have to partially agree with the first objection, at least in the sense that people’s beliefs about such groups are not quite the same as beliefs about demographic groups. All groups are different from all other groups, in some ways, and people do not think in absolutely the same way about all groups. People realize that “the girls on a soccer team” is a smaller group than, say, “people in China.” They realize that there is an element of choice in becoming a Democrat or joining a bowling league, whereas there is little or no choice in deciding one’s race, ethnicity, or gender. And so on.

There are, however, two problems with this attempt to refute the relevance of sports teams and experimental groups from the realm of stereotypes. First, it is not clear that thinking about those groups differs psychologically from other groups that researchers have considered to be stereotypes.

People most likely think about Germans differently than they think about the Yankees, but, then, they also probably think about Germans somewhat differently than they think about Nigerians, lawyers, Hindus, or the immensely wealthy. Thus, it is not clear that there is any psychological rationale for excluding sports teams or experimental groups from the realm of “stereotypes.”

Second, and even more important, however, whether or not distinctions between beliefs about types of groups are justifiable, no scholarship defining stereotypes as inaccurate has ever made such distinctions. No scholarship emphasizing the inaccurate nature of stereotypes has ever included a caveat such as “stereotypes are overgeneralized, invalid, irrational, and rigidly resistant to change, *but only for certain types of groups; many other beliefs about groups are appropriately generalized, valid, rational, and flexible in response to new information.*” Except for Allport (1954/1979, discussed later), I have yet to find anything like the italics that appears anywhere in the social science literature that defines stereotypes as inaccurate!

But here is the kicker. Let’s say the above paragraph is wrong and simply reflects my ignorance.

There is so much scholarship on stereotypes, prejudice, and the various “-isms” (sexism, racism, etc.) that perhaps I have missed the scholarship that has indeed presented a clear theoretical analysis, backed with strong data, demonstrating that certain types of stereotypes are irrational and invalid but others are rational and valid. Great! If some stereotypes are irrational and invalid and others are rational and valid, then one cannot define all stereotypes as inaccurate!

So much for the first objection (that beliefs about sports teams and experimental groups are not really stereotypes). Let’s now consider the second objection that, in my examples of

experimental groups and sports teams, there is clear and objective evidence that one group really is better than or different than another. I reply, "And your point is?" The classic view of stereotypes is rarely qualified by statements such as "Stereotypes are inaccurate, EXCEPT when perceivers have clear and objective evidence of group differences; then stereotypes are often reasonable, accurate, flexible, and nicely in touch with reality." At least I have almost never seen it qualified in this manner.

Allport (1954/1979) readily acknowledged that many people may hold reasonable and rational beliefs associated with social categories, but he did not consider them stereotypes. Although some recent scholarship has begun to take seriously the potential for accuracy in stereotypes (see the next several chapters), the classic view—emphasizing the inaccurate and nasty effects of stereotypes—has rarely, if ever, included qualified conclusions that acknowledge strong potential for accuracy when groups are chosen and when there is objective evidence of group differences (see, e.g., Devine, 1995; Fiske, 1998, 2004; Jones, 1986, 1990; Jost & Banaji, 1994; Stangor, 1995).

Nonetheless, you might truly believe that stereotypes have considerable potential for accuracy when people choose their group memberships or when perceivers have clear, objective information about group differences. If so, then, on this point, you and I are in complete agreement. Let's consider what this means regarding assumptions about the viability of assuming all beliefs about groups are inaccurate.

Do people choose their religion? Adults do, except in societies that oppress people based on religion. Therefore, if you believe that people can be accurate in their beliefs about groups when people choose to enter those groups, then you are compelled to also believe that it is possible that people hold accurate stereotypes regarding Catholics, Lutherans, Jews, Hindus, Muslims, and animists. We have now moved quite quickly from considering it reasonable to consider accuracy in beliefs about the Brazilian national soccer team or experimental groups to considering it reasonable to take seriously the possibility of accuracy among the types of real groups about which real people often hold exactly the type of stereotypes about which social scientists have so frequently expressed dismay and concern.

But wait! Let's return to the situation where societies force religion on people. If you and I both agree that some societies oppress people based on religion, then we are also agreeing that it is reasonable to consider some societies more religiously oppressive than others. Are such beliefs necessarily inaccurate and irrational? If you answered yes, you have now entered a land of paradoxical logical incoherence. ("All beliefs about groups are inaccurate; I believe some societies are religiously oppressive; this belief must be inaccurate—i.e., I willfully hold an inaccurate belief?!!!" Of course, you have no choice but to be incoherent. If you believe that no groups are religiously oppressive, this, too, is a belief about groups and must also be inaccurate!) If you have answered "no" (to the "are all beliefs about groups inaccurate" question), then you have further discovered for yourself that you believe that some stereotypes may be accurate (in this case, stereotypes about the religious oppressiveness of different societies and cultures).

And what about other groups, such as political groups (Republicans, Democrats, liberals, conservatives, socialists, communists, fascists), occupational groups, regional groups, volunteer and professional organizations, etc.? There is a great deal of choice in these group memberships, isn't there? So, if your position is that there may be accuracy when people choose their group memberships, then you are agreeing that many stereotypes (those involving

chosen group memberships) have considerable potential for accuracy (even if you are not ready to agree that all stereotypes have considerable potential for accuracy). Consequently, it cannot possibly be immoral to consider this possibility; nor can it be reasonable to define such beliefs as inherently inaccurate.

Let's return to the other aspect of experiments and sports that some might consider "not relevant" to understanding stereotypes—presence of objective evidence. Census data, standardized test scores, scientific studies, etc., all produce objective evidence about the existence of sex, race, ethnic, regional, and social class differences. So, if your position is that people may be accurate when they have clear objective evidence, then your position inherently assumes that any stereotype for which objective evidence is available could have some, or even high, accuracy. In other words, when people have, use, and understand clear objective evidence about racial groups, ethnic groups, and national groups—exactly the type of groups that have been a central focus of classic social science perspectives on stereotypes—their stereotypes may be accurate. I agree. This is one way to understand why defining stereotypes as inaccurate is unreasonable, and why it cannot possibly be immoral to take stereotype accuracy seriously.

THE ACCEPTANCE OF GROUP DIFFERENCES II: KNOWN GROUPS VALIDITY

"Validity," in the scientific literature, is a close sibling of "accuracy." "Valid" conclusions are those well-justified and believable. "Valid" measurements succeed at measuring what they are supposed to measure. So, if measure X indicates that Fred has high self-esteem, if it is valid, Fred most likely really does have high self-esteem.

Issues of validity can come up anytime, but most often explicitly come up in psychology when researchers develop new measuring instruments. How do we know that Dr. Smith's new self-report scale measuring "motivation to watch TV sitcoms" actually measures motivation to watch sitcoms? It doesn't just because he says so. He needs some sort of evidence. Although a review of all the types of validity evidence is beyond the scope of this chapter, one type in particular is very relevant.

"Known-groups" validity (Cook & Campbell, 1979) refers to validating a new questionnaire by administering it to groups who would be well-known to differ on the measure, if it really measured what it is supposed to measure. For example, let's say we could identify two groups of people: one group who watches 20 hours of sitcoms a week and another who never watches sitcoms. If Dr. Smith's measure is valid, the first group should score higher on Motivation to Watch Sitcoms than the second group.

Validity—one of the core, essential ingredients for any psychological research—*takes for granted that groups differ in many ways* and uses that knowledge in the service of advancing science. To use several examples more relevant to stereotypes than my sitcom one, on average, Whites should show more anti-Black prejudice on all sorts of measures than do African Americans, Catholic priests should score higher on measures of religiosity than do atheists, and conservatives should score higher on measures of right-wing authoritarianism than do liberals. In this context, it is, apparently, not merely moral to believe that groups differ, but assumed and exploited by competent psychological scientists seeking to develop new instruments. If exploiting "known" group differences is part of normal science, then "knowing" that groups differ (i.e., stereotyping them) cannot possibly be inherently immoral or necessarily flawed.

THE ACCEPTANCE OF GROUP DIFFERENCES III: CULTURAL
PSYCHOLOGY AND MULTICULTURALISM

Cultural psychology has become something of a cottage industry of difference documenting. East Asians are more “collectivist” than “individualistic” Western Europeans and Americans (Markus & Kitayama, 1991). East Asians also think in fundamentally different ways than do Westerners (Norenzayan & Nisbett, 2000). Within the United States, a “culture of honor” helps explain both why southerners in the United States are more prone to violence than are northerners and why African Americans (most of whom either live in the south or whose family emigrated from the south) are more prone to violence than are Whites (Nisbett & Cohen, 1996).

In many universities, there is widespread support for “multiculturalism.” Why? Although different proponents have different rationales, themes typically emphasize understanding and respecting the beliefs, values, and practices of people from different groups and backgrounds than oneself. For example, Pinderhughes (1989) titled her book, which is essentially an extended treatise advocating multiculturalism in therapy, *Understanding Race, Ethnicity, and Power*. Presumably, she believes that such understanding can be attained; otherwise, she would not have written the book or given it that title.

This constitutes a call for an increase in the accuracy of people’s beliefs about (understanding of, insight into) those from other cultural backgrounds. I would call that an increase in the accuracy of their stereotypes. Indeed, Pinderhughes (1989, p. 147) makes essentially the same points when she identifies several abilities that enhance people’s multicultural competencies, two of which are:

- “the ability to control, and even change false beliefs, assumptions, and stereotypes . . .”
- the ability to respect and appreciate the values, beliefs, and practices of persons who are culturally different . . .”

On the one hand, she is clearly casting her view along with traditionalists emphasizing the inaccuracy of stereotypes (“... change false ... stereotypes ...”). Apparently, however, she also believes that not all beliefs about groups are inaccurate. Changing false beliefs only makes sense if one is making them less false (more accurate), which, of course, implies the possibility of having and holding accurate beliefs about groups. Respecting others’ values, beliefs, and practices makes sense primarily if one has a reasonably clear (accurate) sense of what those values, beliefs, and practices are.

The discussion of group differences found in cultural psychology and multiculturalism are, unlike sports teams and experimental groups, very much like the types of group stereotypes often railed against by social scientists when it is obvious that “stereotypes” are being discussed. National groups, regional groups, and racial/ethnic groups are all fair game for cultural psychologists and multiculturalists, with one caveat. As long as we are demonstrating how open-minded, tolerant, sensitive, and caring we are, it is permissible, even good, for us to “understand group differences.” So, in contrast to a social problems context, where believing in group differences constitutes lowdown dirty stereotyping, in a (multi-)cultural context, recognizing and being “sensitive” to group differences shows how benevolent and egalitarian we are.

Embracing cultural psychology and multiculturalism, while rejecting stereotypes, may make sociopolitical and rhetorical sense, because it positions the embracer/rejecter as an unbigoted egalitarian who respects others. But, from the scientific standpoint of evaluating the validity of people's beliefs about groups, embracing cultural psychology and multiculturalism while rejecting stereotypes as inherently inaccurate is logically incoherent. When a belief about groups is held by laypeople, many social scientists assume that it is inaccurate and immoral. When a belief about national or ethnic groups is held by social scientists, however, it constitutes cultural psychology, in which case that belief is scientific and validated on the basis of evidence. I can't help but wonder—what do the psychologists who perform and support cultural psychology think about the accuracy of beliefs held by people who read their work and accept its conclusions? Presumably, they want people to believe their work. They are (or, at least, should be) compelled to adopt the position that, therefore, people's beliefs about groups are not necessarily inaccurate.

But if people's beliefs about groups are not necessarily inaccurate, then we could not assume that they are inaccurate even when those beliefs are *not* based on hard research. An absence of research does not render a belief wrong. They may be wrong, but until we do the research, we do not know. This is yet another reason why it should not be immoral to consider the possibility that people's beliefs about groups might, at least sometimes, be accurate.

Part II: Are Stereotypes, By Definition, Inaccurate?

OK, so it is not immoral to consider the possibility that people might accurately perceive groups and group differences. "Merely perceiving group differences," you say, "misses the point of stereotypes. What makes a belief a stereotype," you continue, "is that it is irrational and inaccurate." So, you do not deny the possibility that beliefs about groups may be accurate; you just do not consider those beliefs stereotypes. In fact, you might even see all those other folks who distinguish between, say, stereotypes and multicultural beliefs (e.g., Pinderhughes, 1989) as doing something reasonable, because all they are suggesting is to change the inaccurate stereotypes to more accurate, nonstereotyped beliefs.

To which I respond, "Even though you do not consider all beliefs about groups to be stereotypes, you do, apparently, consider stereotypes inaccurate by definition. If so, therefore, you have two choices, and I suspect you won't like either one:

1. Live with the consequences of your definition, or
2. Change your mind.

There is nothing logically incoherent about defining stereotypes as inaccurate beliefs about groups, as long as one then provides some clear basis for distinguishing stereotypes from accurate beliefs about groups. There are, however, several problems with this: (1) Researchers have almost never stated the criteria they use for distinguishing inaccurate (by definition) "stereotypes" from accurate "nonstereotypes"; (2) instead, they interpret research as if any belief about a group or, indeed, any group label constitutes a "stereotype"; and (3) if stereotypes are defined as inaccurate, then absent evidence of inaccuracy, beliefs cannot be known to be stereotypes. The severe implications of these points are discussed next.

THE LIPPMANN DETOUR: ORIGIN OF THE EMPHASIS ON
STEREOTYPE INACCURACY

Stereotype research started off on a really bad detour, and it wasn't even social scientists' fault. Walter Lippmann (1922/1991), one of the most influential journalists and intellectuals of the early 20th century, was the first person to use the term "stereotype" to refer to people (it previously referred to typesetting in printers). He characterized stereotypes as "pictures in the head" regarding the people in other groups (racial groups, national groups, etc.).

Now, Lippmann was a very intelligent guy working in a profession that was all about words. So I suspect that he chose this term—pictures in the head—very carefully. Think about a picture—literally. They are two-dimensional images of things not moving through space or time. In other words they are:

1. an oversimplification of actual reality and
2. fixed and unchanging.

Metaphorically, this definition nicely captures the older, "classic" view of stereotypes (as shown in this chapter, one to which many scholars still subscribe) as inaccurate because, after all, they oversimplify a complex social reality and are "rigid," fixed and unchanging. Lippmann, though, was a journalist, and social science was in its infancy at the time anyway. So, we can't really blame him. Nonetheless, it is clear that he reached this view of stereotypes without much of what, here and now, would pass for scientific evidence. At the time, social scientists quickly picked up on his views, and this perspective on stereotypes—as oversimplified, inaccurate, and irrationally rigidly resistant to change—dominated perspectives on stereotypes for decades. And, as we shall see, in a slightly more subtle form, it is alive and well today.

LOGICAL/CONCEPTUAL PROBLEMS WITH DEFINING STEREOTYPES
AS THE SUBSET OF BELIEFS ABOUT GROUPS THAT ARE INACCURATE

The first 20 or so pages of this chapter conclusively refuted the (straw?) argument that all beliefs about groups are inaccurate. Therefore, this section considers an alternative way of considering what it means to define stereotypes as inaccurate. Perhaps when some scholars define stereotypes, they do not mean that all beliefs about groups are inaccurate. Instead, perhaps what they mean is that stereotypes are the *subset* of beliefs about groups that are inaccurate. Stereotypes are inaccurate. Accurate beliefs about groups can exist; it's just that (according to this view) they are not stereotypes. So, let's consider the viability of this variation of defining stereotypes as inaccurate.

Why evidence of error and bias in stereotyping does not justify defining stereotypes as inaccurate. One argument for defining stereotypes as inaccurate or unjustified is the abundant evidence that stereotypes are not always accurate and do lead to biases (see Chapters 4 and 5). Such an argument, however, is fatally flawed for two reasons.

First, studies showing inaccuracy or bias in *application* of a stereotype to perceptions regarding an individual target cannot be used as an argument that the stereotype is itself inaccurate. A good tool may be sometimes used inappropriately. That does not make the tool itself bad (with the possible exception of situations where the tool is almost always used for

bad—the issue of whether relying on stereotypes to judge individuals increases or reduces accuracy will have to wait for Chapter 18).

Demonstrations of inaccuracy in use of a stereotype do not invalidate the stereotype itself. Inaccurately guessing that it is warmer in Trenton, New Jersey, than Anchorage, Alaska, today (such a guess can be inaccurate when New Jersey gets hit with an unusual cold spell, or Alaska with a warm spell) does not invalidate the belief that Alaska is usually colder. Similarly, guessing that a student from a lower social class background has done more poorly in school than a student from a middle class background, even if wrong or biased in a particular instance, does not invalidate the belief (indeed, the fact) that, on average, there are social class differences in academic achievement.

Second, the argument that stereotypes are inaccurate because of the abundant research evidence of stereotype bias fails to consider the full range of evidence regarding stereotypes. It is certainly true that stereotypes cause biases. And it is easy to cite lots of studies demonstrating this. Such evidence, however, is highly selective. The big picture would have to include not only the studies that produce bias but also all those showing (1) accuracy and (2) lack of bias. Chapters 16 through 18 review the evidence on accuracy; for now, I simply direct you to the previously reviewed evidence showing that stereotype biases in person perception are, on average, very small (see the meta-analyses summarized in Table 6-1 in Chapter 6).

To look at any small part of a concept or phenomenon and to then define that concept or phenomenon on the basis of that small part is always inappropriate, as two examples readily show.

1. Babe Ruth struck out a lot—indeed, he struck out more often than any other player of his era. By the “one can define a phenomenon by a small part” notion, one would, therefore, be “justified” in concluding that he was a terrible baseball player.
2. People who exercise frequently periodically suffer injuries from falls, collisions, and overuse. By the “one can define a phenomenon by a small part” notion, one would be compelled to conclude that physical exercise is a major health hazard that should be avoided.

In fairness, no scholar has ever made the claim that “you can define a construct or phenomenon on the basis of a small and selective part.” Instead, what happens is more subtle:

1. Researchers selectively focus on the biases produced by stereotypes and only rarely study accuracy, thereby producing a scholarship replete with demonstrations of bias and with only few demonstrations of accuracy.
2. They then cite this evidence as pervasively demonstrating bias (e.g., Aronson, 1999; Devine, 1995; Jones, 1986, 1990; Jost & Kruglanski, 2002), thereby apparently justifying theoretical conclusions emphasizing the erroneous and biased nature of stereotyping.
3. They also systematically ignore, actively deny, or, perhaps, are simply unaware of the accumulating evidence showing that stereotypes are often quite accurate (see Chapters 17 and 18) and that bias is small (see Chapter 6).

Third, bias sometimes increases accuracy (see, e.g., Chapters 10 and 18 and Jussim, 1991). Therefore, one cannot make a knee-jerk leap from “bias” to inaccuracy. Only studies that test for accuracy are relevant to evaluating whether even the stereotype biases that have been demonstrated constitute sources of inaccuracy. For example, consider someone who believes that new graduate students in social psychology, in general, are less expert at research than people who have recently received their PhDs in social psychology. If given little information about two people, one a new PhD, one a new graduate student, one’s “bias” may be to guess that the new PhD has more social psychological expertise. Although there may be some rare exceptions, in general, such a “bias” will increase the accuracy of one’s judgments regarding these two people.

Let’s assume we rig an experiment where there is no truth—say, as social psychologists usually do, by using fictitious targets—and that we find evidence of this PhD “bias.” But is the bias inaccurate? Does it increase or reduce the accuracy of judgments regarding these targets? The demonstration of bias alone, without any test for accuracy, is incapable of demonstrating inaccuracy.

Furthermore, many of the studies most commonly cited as evidence of stereotype bias did not even test for accuracy! (See, e.g., nearly all of the studies reviewed in any of the major reviews of stereotypes and person perception, such as Brewer, 1988; Dovidio & Gaertner, 2010; Fiske & Neuberg, 1991; Jones, 1986, 1990; Kunda & Thagard, 1996; Neuberg, 1994; Snyder, 1984.) Therefore, none of these demonstrations of “bias” justifies assuming that stereotypes are inaccurate.

Always inaccurate? OK, so studies demonstrating bias do not justify defining stereotypes as inaccurate. Still, we are allowed to define things pretty much as we please, as long as we stick to our definitions. So, perhaps one can still define stereotypes as inaccurate, as long as one is sure that the term “stereotype” only refers to inaccurate beliefs. Doing so requires that stereotypes must *always* be inaccurate. For example, if we define the flu as an illness caused by a viral infection, then all flus must be caused by viruses. Given this definition, fever, cold symptoms, nausea, and fatigue produced by something other than a virus is not the flu (maybe it is an allergy or bacterial infection). If we define an active volcano as one that is spewing fumes, gas, ash, or lava, then all active volcanoes must be spewing fumes, gas, ash, or lava. No exceptions.

On logical grounds, therefore, definitions of stereotypes that presumed inaccuracy included a tautology that limited their utility. If all stereotypes are inaccurate by definition, then only inaccurate beliefs about groups can be considered stereotypes. *Accurate* beliefs about groups, then, must constitute a different phenomenon altogether.

This is not necessarily a logical problem as long as we stick to our definition and live with its implications. It is logically possible to consider only inaccurate beliefs about groups to be stereotypes. However, defining stereotypes in this manner does create serious problems with respect to interpreting existing research on “stereotypes” and conducting new research. The problem can be characterized in a simple logical form.

If stereotypes are defined as inaccurate beliefs about groups:

1. accurate beliefs about groups are not stereotypes, and
2. beliefs of unknown validity cannot be known to be stereotypes.

The core problem with this perspective is that it sets the standard for figuring out whether some belief is a “stereotype” exorbitantly high. For this perspective, only when beliefs have

been empirically demonstrated to be inaccurate can one conclude that they are “stereotypes.” Absent evidence of inaccuracy, a belief about a group cannot be known to be a stereotype. The consequence of this variation of stereotype inaccuracy is to invalidate nearly all existing research on “stereotypes.” Because so few studies have actually demonstrated that the beliefs under study are inaccurate, within the context of this interpretation, they could not be known to be “stereotypes.”

Case study: Taking Allport's (1954/1979) definition seriously and applying it to social science research on stereotypes. If I were to define stereotypes as inaccurate, I would have used Allport's (1954/1979) definition (“an exaggerated belief associated with a category”). It is by far the clearest and most logically coherent because he specifically stated what is and is not a stereotype. He specifically excludes accurate beliefs about groups from his concept of stereotype. He does not deny that people may have accurate beliefs about groups—he just does not consider such beliefs stereotypes.

So, in order to study Allportian stereotypes as, say, in a social psychological experiment, one would have to first (1) identify a group of people who exaggerated real differences between one or more groups and then (2) conduct one's investigation into effects of this stereotype on people's perceptions and judgments.

No research in which stereotypes have been defined as inaccurate has ever met this standard. Why? Because social psychology's premature dismissal of accuracy (see Chapter 10) long prevented it from even seeking evidence about whether particular beliefs about groups exaggerate real differences. In other words, in order to know whether a belief is an Allportian stereotype, one has to:

1. first assess people's beliefs about groups,
2. then compare those beliefs to appropriate criteria, and then
3. only if one found that people exaggerated real differences could one conclude that one was studying a stereotype. If the evidence indicated that people's beliefs were either accurate or underestimated real differences, one would be compelled to conclude that the belief was *not* an Allportian stereotype. And if one had not collected any data on people's beliefs, one could never know whether or not a belief was a stereotype!

No research on “stereotypes” has ever been framed as follows:

“Is this belief about that group a stereotype? We are going to figure out whether THIS belief about THAT group is a stereotype by assessing whether that belief exaggerates real differences. If THIS belief does exaggerate differences, we will conclude that it is a stereotype. If THIS belief underestimates or accurately approximates real differences, we will conclude that it is not a stereotype.”

Unfortunately, that is precisely how the question *must* be framed and answered before one can know one is studying an Allportian stereotype. If that question is not answered *prior* to conducting a study on “stereotypes,” one cannot know that one is actually studying a stereotype! So, taking this definition seriously requires a very specific order of research operations, if one wishes to study effects of stereotypes on memory, processing, attributions, evaluations, etc.: (1) One must first demonstrate that people hold a belief about groups that exaggerates real differences and then (2) one can study effects of that stereotype.

This problem characterizes *any* definition of stereotypes as inaccurate, not just Allport's definition of stereotypes as exaggerations. No matter how one defined "inaccurate" (which, in addition to exaggerations, could include inventing nonexistent differences, reversing real differences, or underestimating real differences), one would still need to empirically demonstrate that some belief about a group was inaccurate before one could consider it to be a stereotype.

Instead, in practice, that first crucial step of demonstrating inaccuracy is simply ignored. Did Katz and Braly (1933) demonstrate that people's beliefs about U.S. racial and ethnic groups exaggerated real differences before characterizing them as stereotypes? Did Snyder, Tanke, and Berscheid (1977) first demonstrate that beliefs about physical attractiveness invent or exaggerate real differences *before* studying self-fulfilling effects of stereotypes? Did Taylor, Fiske, Etcoff, and Ruderman (1978) demonstrate that beliefs about sex differences exaggerate real differences before studying their effects on memory? Did Pinderhughes (1989) demonstrate that people exaggerate ethnic, racial, and cultural differences before characterizing lay beliefs as stereotypes? Did Krueger and Rothbart (1988) and Locksley, Borgida, Brekke, and Hepburn (1980) show that people's beliefs about men's and women's assertiveness exaggerate real differences *before* examining effects "sex stereotypes" on person perception? Did Cohen (1981) show that people's beliefs about librarians and waitresses exaggerate real differences before studying effects on memory?

In all of these cases, the answer is a clear "no"—and the list goes on and on (see, e.g., *all* [not any—all] of the studies cited as addressing stereotypes reviewed in Aronson, 1999; Brewer, 1988; Fiske, 1998, 2004; Fiske & Neuberg, 1991; Kunda & Thagard, 1996; McCrae, Stangor, & Hewstone, 1996; Nelson, 2002; Oakes, Haslam, & Turner, 1994). Now, in fairness, many of these researchers themselves *did not* define stereotypes as inaccurate, so they are not being logically incoherent. But that is irrelevant to my point here. *If* one does define stereotypes as inaccurate, one is logically compelled to exclude nearly all social scientific research framed as examining "stereotypes" because so little of it has first demonstrated that people's beliefs are inaccurate.

In other words, holding social psychology to Allport's definition—or *any* definition that assumes stereotypes are the subset of beliefs about groups that are inaccurate and that other beliefs can be accurate but they are not stereotypes—would mean concluding that decades and decades of research framed as addressing stereotypes really has not. None first demonstrated inaccuracy, so none could be sure of having studied a "stereotype" (by this definition). Indeed, there would be nothing left—no studies of the role of stereotypes in expectancy effects, self-fulfilling prophecies, person perception, subtyping, memory, etc. Poof. We would have to throw out the baby, the bathwater, the tub, and the bathroom, and indeed tear down the entire scientific and empirical house on which all our current understanding of "stereotypes" exists, because almost none of it is known to address effects of exaggerated or erroneous beliefs about groups. Such are the consequences of taking this definition seriously.

(And now a brief but important tangent. This situation is exacerbated by those arguing that stereotype accuracy cannot or should not be assessed [e.g., Fiske, 1998, 2004; Stangor, 1995]. If stereotype accuracy can never be assessed, it will be impossible—completely and utterly impossible—to demonstrate that a belief about a group is inaccurate. In this event, "stereotype" as an inaccurate belief or exaggeration of real differences becomes an empty set—because [in]accuracy can never be assessed, we can never know any belief is a stereotype.)

In the sciences, we are stuck with our definitions. If we define a bird as an animal with feathers that flies, then we are required to consider penguins and ostriches something other than birds. And if we are not sure that a chicken can fly, we cannot be sure that it is a bird—unless we recognize that our definition is wrong.

I do not mean to single out Allport (1954/1979) for criticism. Instead, I do mean to flesh out the implications of taking seriously defining stereotypes as inaccurate, of which Allport is only one example. Essentially the same analysis applies to almost any of the vast social science scholarship that defines stereotypes as inaccurate.

Perhaps the worst problem with defining stereotypes as inaccurate is that it would mean that the vast knowledge gained by the social sciences about how people think about groups and group differences cannot be considered to provide any insight into “stereotypes.” Taken seriously, this cost of defining stereotypes as inaccurate is, for me, simply too high and, I think, too high for anyone who believes that the social sciences have indeed provided a great deal of important and useful information about stereotypes.

There is, however, a second, almost equally important problem with defining stereotypes as inaccurate. This is discussed next.

Empirical and conceptual mushiness. Beliefs rarely fall neatly into the categories “accurate” and “inaccurate.” As discussed in Chapter 10, accuracy is quantitative, probabilistic, and a matter of degree; it is not absolute. This means, therefore, that vanishingly few beliefs will be perfectly accurate or perfectly inaccurate. If most beliefs are some mix of accurate and inaccurate, what does it mean to define a stereotype as inaccurate? At what point does a belief become sufficiently inaccurate to justify considering it a “stereotype”? 99.999% accurate? 95% accurate? 90% accurate? 70% accurate? I know of no scholarship that has ever defined stereotypes as inaccurate that has also identified the point at which a belief crosses over from being an “accurate” belief about a group to being a “stereotype.”

I suppose it could be done, but it has never been done. Why not? The most benevolent explanation is that, even today, many researchers do not recognize, or simply do not explicitly acknowledge in their scholarship, that when they demonstrate some sort of psychological process that appears to be imperfect, biased, or error-prone, it usually provides no direct information whatsoever—none, zero, nada—about the (in)accuracy of people’s beliefs (see Chapter 10). Therefore, some researchers may truly believe that the error and bias research shows that stereotypes are inaccurate. If we already have the “answer” (stereotypes are inaccurate), there appears to be no reason to set a cutpoint for (in)accuracy.

Furthermore, many people who do not actually study accuracy often seem to consider accuracy an all-or-none phenomenon, whereby any degree of inaccuracy, any imperfection, means that people are “inaccurate” (see Jussim, Harber, Crawford, Cain, & Cohen, 2005, for a review). This seems to be the implicit logic of those who rely on the error and bias research to claim that people are frequently inaccurate.⁶ There are lots of demonstrations of error and bias, but they almost never compare degree of error and bias to degree of accuracy, because they almost never assess accuracy (again, see Chapter 10). By virtue of never assessing accuracy, the issue of how wrong people have to be before we consider their beliefs about groups to be stereotypes simply never comes up. This means, however, that we have no standard for deciding whether any particular belief is a stereotype.

In practice, this mushiness has allowed researchers who define stereotypes as inaccurate to interpret any and every group belief or label as a stereotype. Scientifically, however, absent a

standard for accuracy, this conceptual mushiness means that we cannot know whether any belief is a (defined as inaccurate) stereotype. If we have no basis for knowing whether any belief is a stereotype, then, again, we have no basis for interpreting any of the research framed as addressing “stereotypes” as actually doing so.

ON THE POLITICAL/RHETORICAL/PSYCHOLOGICAL BENEFITS OF SUCH A PROBLEMATIC DEFINITION

One might wonder why, therefore, if there are so many problems with defining stereotypes as inaccurate, doing so has endured so long and has so heavily pervaded our culture. There is no hard evidence on this question, but the socio-psychological/political benefits of promoting such a view should be apparent:

1. The most benevolent interpretation is that people’s sincere and genuine concern with curing the ills of the world has led them to focus on the ills caused by stereotypes. These concerns, however well-intentioned, have the unfortunate side effect of producing a profoundly distorted image of stereotypes in particular and people’s thought processes in general. If the social sciences (and broader society) focus entirely on the ills, they risk seeing nothing but ills.
2. A somewhat less benevolent interpretation is that few researchers actually have studied accuracy, and many commit many of the logical errors or subscribe to many of the myths and fallacies discussed in Chapters 10 through 12. If one makes false assumptions (e.g., “the research on bias demonstrates pervasive inaccuracy”), then one may make the honest mistake of believing one is justified in defining stereotypes as inaccurate.
3. Those promoting such a view, by implication, position themselves as good, decent egalitarians concerned about defending the weak members of stigmatized groups from the oppressive and hegemonic practices of the powerful perpetuated, in part, through the purveying of pernicious and irrational stereotypes.
4. By defining stereotypes as inaccurate, and simultaneously denying or ignoring the need to collect data on their accuracy, many social scientists get to have their cake and eat it, too. Keeping in mind the overwhelming emphasis on error and bias in the social sciences (see Chapters 4 through 10), they get to indict laypeople’s beliefs about groups as inherently inaccurate (*by definition*) without ever having to do the hard research necessary to evaluate the accuracy of lay beliefs. By not actually doing research assessing accuracy, they can avoid any possibility of ever obtaining any evidence that people’s beliefs are, in fact, accurate, which, in turn, means that they never risk obtaining evidence that disconfirms the claim that stereotypes are inaccurate. And, to the extent that finding evidence of stereotype accuracy implicates one as a possible “ist” (racist, sexist, etc.), avoiding assessing the accuracy of stereotypes allows one to never risk being implicated as an “ist.”

I subscribe to Popper’s (1959/1968) views to the extent that, to be scientific, a belief has to be at least capable of disconfirmation.⁷ If the belief that stereotypes are inaccurate is not subject to empirical disconfirmation, it is not a scientific belief. Similarly, when lay beliefs about

groups are simply labeled “stereotypes” and an assumption of inaccuracy is implicitly imported without a priori evidence of inaccuracy, again, the belief is not scientific, because it is not subject to empirical disconfirmation.

These, then, may be some of the fundamental reasons such a problematic view of stereotypes pervades the social sciences. Taken together, they are quite powerful, ranging from political benevolence and posturing to ignorance to something akin to convenient self-delusion. In the future, perhaps those social scientists promoting a view of stereotypes as inaccurate will articulate what they would consider evidence that could disconfirm their view of stereotypes as inaccurate, or the criteria they use for classifying some beliefs (the erroneous ones) as stereotypes and others as nonstereotypes. Until that time, however, the current common belief in stereotype inaccuracy appears closer to religion than to science.

INTERLUDE: DISCARDING INACCURACY BY DEFINITION AND THEN REIMPORTING IT

Ostensibly neutral definitions. Many modern researchers at least partially recognize that defining stereotypes as inaccurate creates far more problems—logical, conceptual, and empirical—than it solves. Consequently, some modern researchers do not define stereotypes as inaccurate and, instead, provide neutral definitions—definitions that do not take a stand on issues of (in)accuracy and (ir)rationality. For example:

Myers (2002, p. 328): “Stereotype: A belief about the personal attributes of a group of people.”

Mackie, Hamilton, Susskind, and Roselli (1996, p.42): “We define a stereotype as a cognitive structure containing a perceiver’s knowledge, belief, and expectancies about some human social group.”

Fiske and Taylor (1991, p. 119): “One can think about stereotypes as a particular kind of role schema that organizes people’s expectations about other people who fall into certain social categories.”

Although these definitions leave open the possibility that stereotypes are bad, inaccurate, irrational, etc., they do not inherently require stereotypes to be bad, inaccurate, irrational, etc.

One might assume, therefore, that these definitions imply a widespread recognition by researchers that stereotypes may be accurate or inaccurate, rational or irrational, rigid or flexible, etc., and that this recognition has prompted a scientifically open and balanced assessment of the conditions under which stereotypes produce socially harmful and irrational errors and biases and when they enhance accuracy in socially beneficial ways. Nothing, however, could be further from the truth.

If modern perspectives were truly balanced, acknowledging both accuracy and inaccuracy in stereotypes, what would they look like? They would be qualified, nuanced, and very cautiously and carefully written. They would acknowledge the abundant evidence suggesting reasonableness, rationality, and accuracy in stereotypes (see Lee, Jussim, & McCauley, 1995; Park & Judd, 2005; Ryan, 2002; and Chapters 16 through 18) and, at the same time, point out that, sometimes, stereotypes produce biases, errors, and self-fulfilling prophecies. While pointing out that these latter effects can be important and constitute some of the main

reasons for concern about the role of stereotypes in prejudice and discrimination, balanced perspectives would also be very careful to point out that the effects of individuals' stereotypes often tend to be small and fleeting, and that people's beliefs about groups are often nicely (though rarely perfectly) in touch with reality.

Outside of the small minority of researchers who actually have studied stereotype accuracy, however, such balance almost never appears in the social science literature, even the literature that does not define stereotypes as inaccurate.

The subterranean reimportation of inaccuracy assumptions into ostensibly "neutral" perspectives. After defining stereotypes in a neutral way, many researchers proceed to implicitly reimport the older view of stereotypes as inherently bad, irrational, and inaccurate through the manner in which they study or discuss stereotypes. This includes frequently concluding that almost any pattern of stereotype use is something bad that should be stamped out or "overcome," as is readily apparent from the discourse that appears in each of the chapters and articles cited above for "neutral" definitions.

For example, immediately following his definition, Myers (2002) states that: "Stereotypes are sometimes overgeneralized, inaccurate, and resistant to new information." Denotatively, there is nothing wrong with this statement. It is factually true. However, denotatively, its meaning is identical to the following statement, which appears nowhere: "Stereotypes are sometimes appropriate generalizations, accurate, and flexible in response to new information." What makes the denotative meaning identical is the word "sometimes," which appears in both statements. Connotatively, however, they emphasize and imply very different things about stereotypes.

As far as I can tell, there is no reason not to present some balance, as in "Stereotypes are sometimes overgeneralized, inaccurate, and resistant to new information, *and at other times, are appropriately generalized, are accurate, and readily change in response to new information.*" Such a statement is both true and balanced.

Or, consider the subtitles of the sections of a chapter in Mackie et al. (1996): "Correspondence Bias in Forming Group Representations" and "Illusory Correlation and Differential Perceptions of Groups" and "Affective [i.e., non-rational] Mechanisms in Stereotype Formation" and "Stereotypes as Justifications for the Status Quo." One could imagine sections titled "Reality as a Source of Stereotypes" and "Rational Mechanisms in Stereotype Formation." But they do not appear.

Or, consider this classic phrase from the same paragraph in which Fiske and Taylor (1991) provided their definition of stereotypes as role schemas: "As we shall see, stereotypes are 'nouns that cut slices'; they are the cognitive culprits in prejudice and discrimination." Although the role of stereotypes in prejudice and discrimination is, at best, unclear (e.g., Park & Judd, 2005), such statements could be dramatically improved merely by adding some balance. For example, "Although stereotypes are sometimes the cognitive culprits in prejudice and discrimination, at other times, they are simply reasonable, rational, and valid perceptions of groups and their individual members."

Again, I do not mean to pick on these particular researchers, all of whom have made important contributions either to understanding intergroup relations or to undergraduate education (Myers's quote is from his textbook, which I [mostly] love and have used for years; and Fiske & Taylor [1991] is a classic). My only point is that this extraordinary lack of balance whereby the dark side of stereotypes is almost exclusively emphasized characterizes nearly all

of the scholarship presented in ostensibly neutral perspectives on stereotypes (see, e.g., American Psychological Association, 1991; Darley & Gross, 1983; Devine, 1995; Jones, 1986, 1990; Plous, 2003; Wilder, 1986).

This state of affairs—an ostensibly neutral perspective into which an overwhelming emphasis on inaccuracy is imported—is nicely captured by Copus's (2005, p. 33) critical review of social psychological perspectives on stereotypes:

Indeed, most current researchers in their heart of hearts . . . believe . . . that stereotypes are false, negative, irrational and pernicious generalizations about a group—generalizations that inevitably foster harmful prejudice and discrimination. While researchers routinely pay lip service to the formal, neutral definition, in practice, apparently, they really believe that stereotypes are inherently evil.

Now, Copus is a lawyer, not a social psychologist; but he is a lawyer who has reviewed the social psychological scholarship on stereotypes to evaluate its implications for legal issues. Why would a lawyer come away with such an interpretation of the social psychological scholarship? For several good reasons:

1. Prejudice, discrimination, and stereotypes are often discussed together, in which the assumption that stereotypes either cause or justify prejudice and discrimination is either implied or explicitly stated (Devine, 1995; Fiske, 1998, 2004; Jost & Banaji, 1994; Plous, 2003; Snyder & Miene, 1994; Stangor, 1995).
2. Stereotypes are routinely discussed in the scholarly literature and undergraduate texts as things that need to be stopped, prevented, changed, or overcome (Devine, 1989; Fiske & Neuberg, 1991; Myers, 2002; Pinderhughes, 1989; Plous, 2003).
3. The overwhelming majority of research on stereotypes focuses on their negative effects. Stereotype phenomena include “bias,” “self-fulfilling prophecy,” “behavioral confirmation,” “out-group homogeneity” (technical term for “they all look alike to me”), “ultimate attribution error,” “exaggeration of real differences,” and many more (see, e.g., Jussim et al., 1995; Oakes et al., 1994, for reviews).
4. Scholarly perspectives are typically imbalanced in that they usually lack much, and often any, discussion of rationality and accuracy in stereotypes.
5. The scholarly literature is replete with systematic attempts to dismiss work demonstrating accuracy or reasonableness in stereotypes (e.g., Fiske, 1998, 2004; Jones, 1986; Stangor, 1995), and with a systematic overemphasis on their power to produce errors and biases (compare, e.g., Chapters 4 and 5 with Chapters 6 through 9).

And if you consider a lawyer too far outside the social science mainstream and as lacking the expertise to be credible, consider this from Schneider's (2004, p. 19) comprehensive book on stereotypes: “When one reads the literature on stereotypes, one cannot avoid the conclusion that stereotypes are generalizations gone rotten” (for those not “in the know”: Schneider is an influential social psychologist who, among other accomplishments, wrote “the book” on person perception [Schneider, Hastorf, & Ellsworth, 1979]).

THE SUPPOSED “BENEFITS” OF STEREOTYPES CAST
PEOPLE AS DUMB AND LAZY

Even when researchers discuss supposed “benefits” of stereotypes, it has typically been in such a manner as to cast people in a negative light. Researchers routinely acknowledge that stereotypes have two “benefits”: they simplify an overly complex world and they are time- and effort-saving cognitive shortcuts (Fiske, 1998; Fiske & Neuberg, 1990; Mackie et al., 1996; Nelson, 2002).

The argument goes something like this. The world is far too complex and rich with information for people to be able to perceive, code, remember, recall, and use all of it. It is just overwhelming. So, stereotypes serve a valuable function by simplifying this overly rich, overly complex world into ways they can be easily stored and used. Similarly, perceiving and remembering every piece of relevant information about a person requires a great deal of effort. Stereotypes, therefore, often function as shortcuts, allowing people to reach judgments about others without exerting all that effort. From the individual’s standpoint, therefore, they perform a valuable effort-saving function.

At first glance, these look like researchers recognizing benefits of stereotypes. But look more closely (or think more deeply) about these “benefits.” There is something . . . not so nice about them, isn’t there?

When researchers “acknowledge” these supposedly positive effects of stereotypes, what are they saying about people? First, they seem to be saying that complexity is too hard for most folks to handle. This does not seem like a particularly flattering view of lay intelligence. Second, they are saying that people are too lazy to judge individuals on their merits. This does not sound like high praise, either. And so, even when researchers “acknowledge” positive effects of stereotypes, it is done in the service of perpetuating a singularly unflattering view of human judgment.

Now, it is true that generalizations—all generalizations, not just stereotypes—help people simplify a complex world. Generalization, however, can be viewed as an extraordinary cognitive and intellectual achievement that marks one of the key strengths of human beings, rather than as a reflection of laziness or simplicity. People who cannot reach generalizations and abstractions are seriously cognitively impaired, and scientific theories often require extraordinary leaps from specific instances to general principles (e.g., Newton’s apple and the law of gravity; Darwin’s finches and evolution). Without the power of generalization, it is unlikely that *Homo sapiens* would have reached their current position as the dominant species on Earth. Why, then, should this extraordinary skill, widely recognized as such in other contexts, all of a sudden reflect simplicity or laziness when people think about groups? Again, the proponents of this view have never provided an answer to this question. The published scholarship provides little or no evidence that many have even considered it.

Overall, therefore, even though some modern definitions of stereotypes appear and, indeed, are neutral with respect to the accuracy and morality of stereotypes, the overwhelming majority of scholarship on stereotypes implicitly or explicitly assumes that stereotypes are inherently immoral and invalid. Making this point implicit rather than explicit in the definition, however, only serves to further distort social science perspectives on stereotypes. Pretending that we define stereotypes neutrally and then (wink wink nod nod) focusing exclusively on their invalidity and negative effects at best produces a highly limited scientific

view of stereotypes (comparable to focusing on Babe Ruth's strikeouts); at worst, it makes us look like politicians or preachers whose preordained conclusions drive our scholarship rather than as scientists whose research drives our conclusions.

Part III: How to Define Stereotypes

After having demonstrated the incoherence of considering all beliefs about groups to be inaccurate, after showing that there is nothing immoral about considering the possibility that some aspects of some stereotypes may have some degree of accuracy, and after showing that defining stereotypes as the subset of beliefs about groups that are inaccurate creates far more problems than it solves, I am ready to provide my definition of stereotype. Almost.

MY FAVORITE DEFINITION (THOUGH IT IS NOT SCIENTIFICALLY TENABLE)

One way to solve this problem is to acknowledge it—as did Brigham (1971) in his comprehensive review of ethnic stereotype research to that point. Brigham pointed out that most researchers and laypeople routinely assumed that stereotypes were invalid, irrational, rigidly resistant to change, etc., despite an almost complete lack of evidence demonstrating invalidity, irrationality, etc. So, he came up with the following definition (p. 31): “An ethnic stereotype is a generalization made about an ethnic group, concerning a trait attribution, which is considered to be unjustified by an observer.”

This is my favorite of all definitions (and it need not be restricted to ethnic stereotypes), because it exquisitely and succinctly captures how people—laypeople and scholars—typically use the word “stereotype”: as a damning indictment of *someone else's* beliefs about a group. *My beliefs are reasonable, rational, and appropriate; yours, at least when they differ from mine, are mere stereotypes.* This definition helps us understand why social scientists can perform research documenting cultural, ethnic, social class, or racial differences and, *at the same time*, condemn *other people's* beliefs about group differences as irrational “stereotypes” steeped in bigotry. It helps us understand why proponents of multiculturalism believe that they know the “truth” about which group differences are important to understand and respect and, *at the same time*, rail against other people's inaccurate stereotypes. Indeed, one of the easiest and most effective ways for Person A to derogate and dismiss Person B's claims about a group is to say, “But that is just a stereotype.” Person B, now implicitly accused of being an “ist” of some sort, is most likely to just shut up and go away, and even if he or she doesn't, has been discredited anyway.

So, as a description of the phenomenology of the use of the word “stereotype,” this is a perfect definition that, in a sentence, captures what I have been writing about throughout this chapter.

Unfortunately, however, it fails as a scientific definition, precisely because it is purely phenomenological and subjective. Its subjectivity leads it into logical incoherence: By definition, if I think your belief is accurate, then your belief is not a stereotype; but if someone else believes your belief is inaccurate, then your exact same belief is a stereotype. Subjectively, this is possible, but scientifically, it is impossible for something to simultaneously be and not be a stereotype. So, as a scientific definition, as amusingly beautiful as this definition is, it fails.

MY ACTUAL DEFINITION

I like nearly all of the modern, neutral definitions. My favorite of these, and one of the simplest, was that provided by Ashmore and Del Boca (1981, p. 21): “. . . a stereotype is a set of beliefs about the personal attributes of a social group.” This allows for all sorts of possibilities not explicitly stated. Stereotypes may or may not:

1. be accurate and rational;
2. be widely shared;
3. be conscious;
4. be rigid;
5. exaggerate group differences;
6. assume group differences are essential or biological;
7. cause or reflect prejudice; and
8. cause biases and self-fulfilling prophecies.

To me, it is good that this definition does not specify these things. That leaves them open for empirical investigation. Sometimes, a stereotype may be accurate and rational; other times it may be inaccurate and irrational. Sometimes stereotypes may be rigidly resistant to change; other times they may be highly flexible in response to social reality. And so on.

TAKING THE NEUTRAL DEFINITION SERIOUSLY

One major point of contention among modern social scientists is the extent to which this neutral definition is taken seriously. Although much scholarship either explicitly defines stereotypes as inaccurate or provides a neutral definition into which inaccuracy, irrationality, lack of justification, etc., is reimported, there are many social psychologists whose writings do take seriously the neutrality of stereotypes, even when they define it somewhat differently (Ashmore & Del Boca, 1981; Judd & Park, 1993; Lee & Ottati, 1995; McCauley & Stitt, 1978; Ryan, 2002; Schneider, 2004; see also most of the research cited in Chapter 17). Therefore, although the views expressed here may be somewhat unusual, they are, in fact, built on a long line of scholarship that dissents from and contests the common emphasis on stereotype inaccuracy.

Stereotypes sometimes are indeed interwoven with prejudice and discrimination. Other times, however, people's beliefs about groups are nicely in touch with reality. One of the great values of truly believing in the neutral definition is that it does not presume that any time a person holds or uses a stereotype, something inherently bad (or good) is happening. Instead, it opens the door for understanding when stereotypes wreak damage, when they simply reflect social reality, and, possibly, when they actually perform a social good.

WHAT IS NOT A STEREOTYPE?

My definition excludes all beliefs about things other than human groups. It excludes beliefs about nonliving things (rocks, houses), about nonhuman forms of life (plants, livestock), and about nonmaterial abstractions (“freedom”). It is possible that there really are no

fundamental psychological differences between these types of beliefs and beliefs about human groups. If so, then the next several decades of research may demonstrate that lack of difference, and, at that time, it might be appropriate to eliminate any distinction between beliefs about human groups and other types of beliefs or generalizations. For now, however, I do distinguish between beliefs about groups of people and other types of beliefs.

My definition also excludes beliefs about individuals. I do not consider David's belief that his wife is 61 inches tall, or Shalonda's belief that her daughter is extroverted, or James' belief that Frank is dependable to be stereotypes. It is possible that stereotypes influence those beliefs, but that is a separate question. And if we assume that A might cause B (stereotype might cause perception of an individual), we are also logically compelled to conclude that A and B (stereotype and perceptions of an individual) are different constructs. Things do not cause themselves.

I urge my colleagues in the social sciences, especially those who either define stereotypes as inaccurate or who emphasize their inaccuracy, to similarly state what sort of beliefs they consider not to be stereotypes and how one would ever know which is which. Then and only then can the logical incoherence, confusion, self-delusion, and hypocrisy that have characterized definitions of stereotypes begin to be rectified.

Epilogue

SUPPORT FOR EGALITARIAN MOVEMENTS

Stereotypes generally emphasize ways in which groups differ. Sometimes, stereotypes reflect and reinforce unequal status and role relationships between groups. In such situations, activists fighting inequality and restriction of freedom are likely to rail against the unfair, unjustified, and inaccurate nature of stereotypes. The individuals who are at the forefront of such movements, and who seek access to previously denied roles and opportunities, are likely to be particularly hostile to any notion of stereotype accuracy (see also Eagly & Diekmann, 2005, for a fuller elaboration of this analysis).

To the extent that many social scientists see their research as serving the goals of increasing equal opportunity and access, many may remain hostile to the notion of stereotype accuracy. Equal opportunities, rights, and access to roles and opportunities are unequivocally good things. Nothing in this book claims or implies otherwise.

Nonetheless, Eagly and Diekmann (2005, p. 30) described the scientific versus political state of affairs quite well: "Because activists rail against the stereotypes that have characterized their groups on the basis of their traditional social position, theorists such as Allport have *overaccommodated by defining stereotypes as necessarily inaccurate*" (emphasis mine).

I would add that this characterization goes well beyond Allport and includes many modern social scientists. Lofty political purposes do not justify faulty scientific claims.

STICKING TO THE TERM "STEREOTYPE"

I have long considered trying to come up with some term other than "stereotype" to refer to people's beliefs about groups. "Stereotype" is heavily loaded with pejorative connotations and widely assumed to refer to irrational, bigoted prejudices. Therefore, it often feels like a

hopelessly quixotic task to convince many people to take seriously the idea that not all stereotypes are inaccurate. I considered jettisoning “stereotype” and just using “beliefs about groups.” I considered keeping the term “stereotypes,” defining them much like Allport (exaggerated beliefs about groups), and then inventing a second term to refer to accurate beliefs about groups (“accutypes”?).

But even if convincing many folks that stereotypes are not inherently bad is tilting at windmills, I decided to stick with the term for several reasons. First, any new term, by virtue of being new, would be cut off from nearly a century of scholarship on stereotypes. People reading about “accutypes” would not necessarily realize that all that work by Allport, Ashmore, Brewer, Brigham, Campbell, Devine, Fiske, Hamilton, E. E. Jones, Katz, LaPiere, Oakes, Schneider, Snyder, Wilder, and a host of others was all addressing the same general topic. And, although I believe much of that scholarship has greatly and unjustifiably overemphasized the bad and irrational in stereotypes, if one looks at the actual studies, they are usually quite good and tell us quite a lot about stereotypes and prejudice (even if what they tell us is that the role of stereotypes in prejudice is much less than once believed—e.g., Park & Judd, 2005).

Furthermore, it may be a hopeless task, but I, as a social psychologist deeply committed to the scientific analysis of human nature and behavior, feel that, at minimum, my field needs to be intellectually honest—both with itself and with the outside world. And, if after 90 years of proclaiming the evils of stereotypes, of proclaiming them to be necessarily inaccurate, unjustified, exaggerated, steeped in prejudice, and so on, we ultimately realize that neither logic, nor empirical evidence, nor the manner in which we have conducted our research supports that view, we cannot just say “never mind.” If it was important to emphasize the inaccuracy of stereotypes for 90 years, and if we discover that perhaps, just perhaps, that emphasis was misplaced, it cannot possibly be appropriate to just move on and ignore nearly a century of misguided conclusions and claims. It behooves us to undo the erroneously dark image of human social thought that we have perpetrated all these decades and replace it with one that is more appropriate to the evidence.

It has now taken me a whole chapter to argue that it is reasonable to take seriously the idea that stereotypes are not always the immoral, invalid, irrational “culprits in prejudice and discrimination” that they are usually cracked up to be. That, however, does not provide a shred of evidence that stereotypes can actually be accurate or reasonable. I do hope, however, that this chapter has presented the logical and conceptual bases for taking seriously the possibility—just the possibility—that not all stereotypes are completely inaccurate and irrational. If so, then empirically and scientifically examining the accuracy of stereotypes becomes an important and open question—rather than one whose importance we have eliminated by definition. In this spirit, therefore, the next several chapters examine the evidence regarding the accuracy of stereotypes.

Notes

1. The correct answers are:

- 1) Men
- 2) African Americans

- 3) Conservative
- 4) Asians, Whites, African Americans (Yes, Asians really have earned higher household incomes than Whites, and have done so at least since the 1990 U.S. Census; they also complete college at much higher rates than do Whites. It is enough to make one wonder about claims about a society structured to protect and perpetuate the Euro-centric and hegemonic interests of a White ruling class . . .)
- 5) Egyptian/Israeli
- 6) Jewish
- 7) Japan-collectivist; Britain-individualist

2. Some of these appear in print, some have occurred at a conference, and one was in a review of a manuscript submitted for publication. At the May 2004 American Psychological Society conference panel on “Stereotyping, Discrimination, and the Law,” “Nonsense” was Lee Ross’s characterization of my description of Brodt and Ross (1998) as showing that relying on an accurate stereotype can increase accuracy of person perception (he is the Ross on that study, which is discussed in detail in Chapter 18 and is readily available to the general scholarly public because it was published in a widely circulated journal). Living “in a world where all stereotypes are accurate . . .” was Susan Fiske’s introductory comment as she began her talk at the same conference. “Disagreeing with civil rights . . .” is also from Fiske (*Handbook of Social Psychology* chapter, 1998, p. 381) and refers specifically to McCauley, Jussim, and Lee’s (1995) concluding chapter to their book, *Stereotype Accuracy* (in that chapter we argued that, in the absence of perfectly diagnostic individuating information, people would make more accurate person perception judgments if they relied on rather than ignored accurate stereotypes—exactly the result empirically found by Brodt and Ross [1998]). Stangor (1995) did not specifically accuse any particular person of “supporting bigots”; instead, he indicted the entire scientific attempt to assess the accuracy of stereotypes as potentially supporting bigotry. In 1990, I submitted an article to *Psychological Review* that argued that, if social psychologists wanted to make claims about the inaccuracy of stereotypes (which, given the frequency of such claims, they apparently wanted to do very much), it behooved them to perform research that actually empirically assessed the accuracy of stereotypes. A reviewer of that draft responded to that section by asking, sarcastically, “What should we be doing, articles with titles like ‘Are Blacks Really Lazy?’ and ‘Are Jews Really Cheap?’” (I took that section out; the article was eventually published by *Psychological Review* [Jussim, 1991]). Nonetheless, that call for research on stereotype accuracy appeared in many other papers (e.g., Jussim, 1990; Jussim, McCauley, & Lee, 1995; Jussim, Eccles, & Madon, 1996) and, in fact, has been answered by many researchers over the last 20 years. This and the next two chapters review that evidence. It is, perhaps, worth noting that, of the scores of empirical studies and meta-analyses reviewed, not a single one is titled anything like “Are Blacks Really Lazy?” or “Are Jews Really Cheap?”

3. Charter schools are a relatively recent innovation. They typically are public schools whose mission and methods have been designed by some community group, rather than by the existing educational administration or local school boards. They are often created because some community group believes the current public schools do not adequately serve a particular group of students or because the community group holds a very different philosophy of education than is practiced in the public schools.

4. It is unlikely that Roxbury Prep founders and administrators would describe themselves as using stereotypes to inform the founding and mission of the school. This is because “everyone

knows” that stereotypes are cognitive evil-doers employed by bigots to oppress and exploit, and I am sure that they would not want to be seen that way. From my standpoint, however, what they did, in part, required the use of accurate stereotypes, whether they would describe it that way or not. Later in this section on morality I explicitly discuss conditions under which it is socially acceptable versus unacceptable for social discourse to recognize group differences.

5. Of course, such research has also done considerable good, too, including contributing to landmark antidiscrimination Supreme Court cases (*Brown v. Board of Education*; *Hopkins v. Price-Waterhouse*) and a slew of less high-profile beneficial outcomes (Aronson, 1999). I am not condemning traditional research on stereotypes, prejudice, and discrimination here; my goal is far more narrow—to refute the specific claim that research on stereotype accuracy should not be conducted because it supposedly causes social harm.

6. This is implicit, rather than explicit. I know of no research that has ever bluntly stated something like “any evidence of error or bias equals inaccuracy.” Researchers do, however, routinely interpret any evidence of error and bias as equivalent to inaccuracy, even in the absence of tests of accuracy (e.g., Darley & Fazio, 1980; Jones, 1986, 1990; Jost & Kruglanski, 2002; see also Krueger & Funder, 2004, for a review).

7. Actually, Popper claimed that scientists should seek to disconfirm their theories. I am not sure that I would go that far. I do think, however, that beliefs, hypotheses, or theories that are not capable of being disconfirmed are not scientific. If we shed this criterion for distinguishing scientific from nonscientific beliefs, then there ceases to be much basis for considering beliefs in Zeus, ghosts, or reincarnation to be unscientific.

16 What Constitutes Evidence of Stereotype Accuracy?

CHAPTER 15 EMPHASIZED the unreasonableness of defining stereotypes as inaccurate. Of course, that does not necessarily make them accurate. Perhaps stereotypes are in fact largely inaccurate. Once the accuracy of stereotypes is no longer decided by fiat (definition), it then becomes a scientific question, not a moral, religious, political, ideological, or philosophical question. It is a question that can and should be answered, not by our social goals, opinions, or belief systems (egalitarian or otherwise), but by the data. How well do people's beliefs about groups, and differences between groups, correspond to what those groups and their differences are actually like?

Chapter 15 started off with a stereotype accuracy “test.” OK, now I can admit it. That test was rigged to be very easy. Why? Because (1) I wished for you to see for yourself that an absolutist claim requiring all stereotypes—all beliefs about groups—to be inaccurate was not viable because it is obvious that lots of beliefs about groups are accurate and, (2) I wanted to demonstrate the need to take seriously a neutral definition of stereotypes—one that does not assume that stereotypes are bad, irrational, immoral, and inaccurate. Researchers defining stereotypes as inaccurate almost never qualify their definitions with statements such as, “Stereotypes only refer to things that are difficult to know about groups; people often know many things about groups that are easy to learn, so we do not consider those to be stereotypes.” Providing an easy “test” was necessary to demonstrate why blanket condemnations of stereotypes as inaccurate are themselves unjustified in an intuitively simple manner.

That test, however, did *not* refute the idea that stereotypes—especially stereotypes about the traits, achievements, and behaviors of various demographic groups—are frequently inaccurate. It would still be unjustified to *define* stereotypes as inaccurate if they were

occasionally accurate but mostly inaccurate. Thus, Chapter 15 on defining stereotypes was not intended to address whether stereotypes are generally accurate or inaccurate.

Hold on to your hats. It will take four chapters, this and the next three, to fully address the scientific research examining the accuracy of stereotypes. This is necessary for the following reasons. First, these chapters reach a very controversial conclusion—that scientific research evidence pervasively demonstrates extraordinary levels of accuracy in social stereotypes. Such a controversial conclusion cannot be reached on the basis of a small handful of studies; it needs to be based on a thorough review of studies on the topic. Second, the studies should be presented in sufficient detail that you, gentle reader, can reach your own conclusions about their results; you should not have to rely on a brief gloss. Third, beyond the complex issues in assessing accuracy reviewed in Chapters 10 through 12, there are additional issues involved in the assessment of stereotype accuracy. These must be reviewed first, before the studies can be understood. Last, because the issue is so controversial, these chapters are peppered with discussions of limitations to and qualifications on the evidence. I want the evidence to be understood for what it does and does not show.

Stereotype Accuracy and Levels of Analysis

A common reaction to research demonstrating stereotype accuracy is “yes, but . . .” A “yes, but” occurs whenever incontrovertible evidence of stereotype accuracy is presented and a person steeped in traditions viewing stereotypes as unmitigated evils responds with “Yes, but what about _____?” (You can fill in the blank with your preferred objection to stereotype accuracy research, if you have one; if not, suffice it to say that some common “yes, buts” include “Yes, but what about self-fulfilling prophecies?” “Yes, but what about biased evaluations?” “Yes, but what about information-seeking biases?” and “Yes, but what about stereotype threat?”). Of course, each “yes, but” can be addressed on its own merits. The discussion of stereotype threat in Chapter 1 was essentially a refutation of a “yes, but,” and Chapters 6 through 9 should refute most of the “yes, buts” arising out of all the expectancy confirmation literature addressed in Chapters 4 and 5.

In general, people engage in the “yes, but” tactic in response to evidence of stereotype accuracy when they (1) are committed to a position emphasizing stereotype inaccuracy, bias, or irrationality, or (2) are unwilling or unable to refute the evidence demonstrating rationality, reasonableness, or accuracy, so that (3) they (perhaps somewhat defensively) attempt to “limit the damage” by acknowledging the existence of some degree of stereotype accuracy (this is the “yes” in “yes, but . . .”) and then returning as quickly as possible to the evidence of bias with which they are generally much more familiar or comfortable (this is the “but” in “yes, but . . .”).

Some “yes, buts,” however, seem so reasonable and are so widely believed to negate any possibility of stereotype accuracy that it is worth spending some time on them. One such common “yes, but” is the following: “Yes, but even a stereotype that is, in some sense, ‘accurate’ as a description of a group mean will not apply to most members of the stereotyped group, because hardly anyone falls on the mean. Therefore, even such stereotypes are inaccurate most of the time.” Let’s examine this more closely. Variations on this “yes, but” summarize a class of criticisms of the notion of stereotype accuracy that has periodically appeared

in the social psychological literature (e.g., Allport, 1954/1979; American Psychological Association, 1991; Fiske, 1998; Hamilton, Sherman, & Ruvolo, 1990; Nelson, 2002; Stangor, 1995):

Even if it can be successfully shown that perceivers accurately judge two groups to differ on some attribute:

1. perceivers cannot assume that their stereotypes of the group automatically fit all members of the group;
2. perceivers cannot apply their belief about the group when judging individuals; and
3. if perceivers do apply their belief about the group when judging individuals, they are likely to be wrong much of the time because few members perfectly fit the stereotype.

If all stereotypes are known to be largely inaccurate (as this logic suggests), the need to assess their accuracy would be rendered moot.

This criticism has some validity, but that validity depends, in part, on what this type of statement means. To the extent that the “perceivers cannot” statements represent moral injunctions rather than statements about accuracy, they are beyond the scope of a consideration of stereotype accuracy (although see Chapter 15 for an extended discussion of the morality of considering the possibility that some stereotypes may sometimes have some degree of accuracy). However, if “perceivers cannot” means “they would reach inaccurate judgments if they did,” these arguments are a central focus of this and the next several chapters.

This line of reasoning’s suggestion, however, that all stereotypes are inaccurate because most members of a group fail to fit a stereotype is only partially justified. It is true that most members of a group will fail to perfectly fit a stereotype. This, however, does not mean that the stereotype is inaccurate. To understand why requires understanding how this reasoning confounds two different levels of analysis and how considerably greater conceptual clarity can be brought to understanding stereotype accuracy by clearly distinguishing among these levels of analysis. Table 16–1 presents an analytic breakdown of different levels of analysis at which accuracy can be assessed.

Stereotypes as perceptions of populations. The first row in Table 16–1 refers to stereotypes: beliefs (or generalizations) about whole populations (typically, but not always, large demographic groups). The level at which one must measure the criterion for assessing the accuracy of beliefs about groups is the population that makes up that group. Claims about the characteristics of New Yorkers (or women or African Americans or librarians) should be compared to the characteristics of a representative sample or the whole population of New Yorkers (or women or African Americans or librarians, respectively). It is not possible to evaluate the accuracy of a belief about Asians in general by using as a criterion the characteristics of my friend Hong. To do so would be equivalent to evaluating the claim that “Alaska is cold” by measuring the temperature at noon on July 4 in Anchorage.

Census figures, results from randomly selected samples, and meta-analyses of hundreds of studies have all been justifiably used as criteria against which to compare the accuracy of people’s stereotypes (discussed later in this chapter and in Chapter 17). Such research,

TABLE 16-1

Identifying the Appropriate Level of Analysis in Studies of Social Perceptual Accuracy		
Level of Analysis	Social Belief Is A:	Level of Criteria for Assessing Accuracy of That Social Belief:
Population	Stereotype Regarding an Entire Population	Population
This level assesses the accuracy of a stereotype about a group. <i>Research examples:</i> Wolsko et al. (2000) McCauley and Stitt (1978) Swim (1994)	<i>Examples:</i> 1. An introductory psychology student believes that White Americans are wealthier than African Americans. 2. A high school teacher believes that teenage boys are better at math than are teenage girls.	1. Income of White Americans and African Americans in a nationally representative sample or in the U.S. Census 2. Meta-analyses of hundreds of studies assessing sex differences in teenagers' math performance
Small Groups	Perception of Differences Between Specific Individual Members of Social Groups	Small Groups
This level assesses the accuracy of beliefs about differences between specific individual targets belonging to different groups. This corresponds to what is frequently termed "stereotypes and person perception." <i>Research examples:</i> Brodt and Ross (1998) Clarke and Campbell (1955) Madon et al. (1998)	<i>Examples:</i> 1. An introductory psychology student sees little difference between the wealth of African American and White students in his class. 2. A high school teacher believes the girls in her class are doing better at math than are the boys in her class.	1. The wealth (net worth; yearly income) of the African American and White students in that student's introductory psychology class 2. Performance in class and on standardized tests of the boys and girls in this teacher's class
Individual	Person Perception	Individuals
This level assesses accuracy in perceptions of individuals, not in perceptions of population or small group differences.	<i>Examples:</i> 1. An introductory psychology student believes that Mary Anne is wealthier than Rashid, who is wealthier than Lois.	1. Mary Anne's, Rashid's, and Lois's wealth

TABLE 16-1

Identifying the Appropriate Level of Analysis in Studies of Social Perceptual Accuracy
(Continued)

Level of Analysis	Social Belief Is A:	Level of Criteria for Assessing Accuracy of That Social Belief:
<i>Research examples:</i>	2. A high school teacher	2. John's, Bonita's, and Lou's
Funder (1987)	believes that John is doing	performance on math tests
Jussim (1989)	better at math than Bonita,	
Kenny (1994)	who is doing better than Lou.	

however, cannot and was never intended to evaluate the accuracy of people's perceptions of individuals from different groups, which requires a level of analysis below that of whole populations.

Small groups: Stereotypes and person perception. The second row of Table 16-1 presents a second level of analysis for assessing accuracy: that of perceptions of differences between individuals belonging to different groups. This is generally referred to in the scholarly literature as "stereotypes and person perception" and is represented by studies that have people rate one or more individuals belonging to different social groups (usually holding their personal characteristics constant—see Chapters 5, 9, and 18).

Let's say a fourth grade teacher assigns higher grades to the girls than to the boys in her class. One might claim that she stereotypes her girls as achieving more highly, but assigning higher grades to girls is itself *not* a claim about whole populations of boys and girls. Whether this is accurate or biased cannot be determined by comparison to the mean achievement of nationally representative samples of fourth grade boys and girls. Instead, determining the accuracy of her higher grading of girls requires comparison of her ratings to (some objective measure of) the achievement of the particular girls and boys in her class (such as a well-validated standardized test). This level of analysis addresses the role of stereotypes in causing systematic inaccuracy in perceivers' judgments about individuals they know personally. Such claims occur at a different, smaller level of analysis than do claims about differences between whole populations.

Assessing the accuracy of the perceived difference at this level of analysis must be accomplished by comparing the perceived mean difference between individual targets from differing groups to the actual mean difference. Research doing so is discussed in Chapter 18.

Person perception. The third row of Table 16-1 presents a third level of analysis: the individual target.¹ At this level of analysis, most stereotype accuracy questions disappear. Accuracy in the perception of differences between individuals belonging to different groups can no longer be assessed. Without some comparison of differences in perceptions of groups (large or small; at minimum, perceivers must evaluate one target from Group A and one from Group B), only accuracy in the judgment of individual targets can be assessed. For example, a business owner's evaluations of employees might correlate .6 with those employees' overall performance, indicating moderately high accuracy. Such accuracy tells us

nothing, however, about whether the owner exaggerates differences between males' and females' job performance.

Stereotype accuracy and level of analysis: Conclusion. Claims suggesting that stereotypes are inaccurate because they do not apply to all individual members of a group (Allport, 1954/1979; American Psychological Association, 1991; Fiske, 1998; Hamilton et al., 1990; Nelson, 2002; Stangor, 1995) are both true and false. The claim that stereotypes cannot possibly apply to all individual members of a group is completely true. The suggestion that this renders stereotypes inaccurate is, however, unjustified because it confounds levels of analysis (population and either small group, individual, or both). A claim about a population cannot be evaluated against the characteristics of an individual, or even small groups of individuals. Consistency between the level of the perception and the level of the criterion must be maintained when assessing accuracy by comparing beliefs about populations (stereotypes) to characteristics of those population groups, and beliefs about differences between small groups of individuals to the actual differences between those small groups of individuals.

Thus, the common “yes, but”—“yes, but stereotypes are inaccurate because they do not apply to all individuals”—completely fails in its attempt to cast all stereotypes as inherently inaccurate because it confounds population and small group levels of analysis.

The one exception: Absolutist stereotypes. Absolutist stereotypes—beliefs that all members of a group have some attribute—will indeed almost always be false, because there are almost always wide variations among individuals. A single exception invalidates an absolutist belief. Just as a belief that the temperature in all locations in Alaska is always below freezing will be disconfirmed by a single reading of 33°F in Juneau on July 15 at 1 p.m., a belief that all Germans are efficient will be disconfirmed by discovery of a single inefficient German.

The vast accumulated empirical evidence on stereotypes, however, has yet to report a single person who holds absolutist stereotypes. Instead, the evidence indicates that most stereotypes are quantitative and probabilistic, not absolute (e.g., Judd et al., 1991; Krueger, 1996; McCauley & Stitt, 1978; Swim, 1994). Probabilistic stereotypes, which permit many exceptions and wide variability, can only be evaluated by comparison to population-level criteria. People who hold absolutist stereotypes undoubtedly exist, and probably make up significant portions of extremist groups such as the Ku Klux Klan and neo-Nazis. Nonetheless, such people are atypical of the participants in most scientific research on stereotypes.

Some Preliminary Evidence That Group Differences Are Broadly Consistent with Stereotypes

Around the world, on average, males are more physically aggressive than females (Brannon, 1999). In the United States, Jews are wealthier than most other ethnic groups; African Americans are more likely to be in jail for committing crimes and more likely to be victims of crime than are others; Asian Americans are more likely to complete college than are others; and people with lower incomes are less well-educated than are people with higher incomes (Marger, 1994; U.S. Census, 2010a,b). These are all verified group differences, and people who believe in them hold more accurate stereotypes than those who do not.

These types of differences seem to fit many stereotypes commonly held. Sometimes they may be held by malevolent bigots; sometimes they may be held by everyday people going about their business; sometimes they may be held by good, decent people interested in equality before the law, who are simply in touch with reality. Accuracy, however, is not assessed by measuring niceness or egalitarian-ness or bigotry. It is assessed by determining the correspondence of belief with reality.

Still, these data are of the “everybody knows” variety, in the sense that “everyone knows” that African Americans are stereotyped as criminals, that Jews are stereotyped as affluent, and that Asian Americans are stereotyped as high academic achievers. I have not, however, presented any evidence that anyone actually believes, for example, that Jews are richer, Asians achieve more highly in school, or African Americans commit and are victimized by more crime. It seems common knowledge that these are widely held stereotypes, but that is not hard scientific evidence. So, the next sections of this chapter review some of the early scientific evidence on stereotype (in)accuracy; the next chapter reviews the modern evidence.

Chapter 2 reviewed some of the earliest evidence supposedly demonstrating inaccuracy in stereotypes and found that evidence to be seriously lacking. Chapter 2 showed that Katz and Braly (1933) inferred inaccuracy from widespread agreement, which was more than a little topsy-turvy, because agreement is usually associated with more, not less, accuracy (see also Chapter 11). It showed that LaPiere’s (1936) study of anti-Armenian prejudice, which was once commonly cited as evidence of inaccurate stereotypes, by modern standards actually provided very little evidence that bore on the accuracy question. And it showed that Hastorf and Cantril’s (1954) classic “they saw a game” study, which has long been viewed as a testament to disagreement and subjectivity, actually provided far more evidence of agreement and objectivity.

Although research on stereotype accuracy did not begin in earnest till the 1990s, the amazing thing is that even the older research, suggestive and inconclusive though it may be, consistently pointed in the direction that, at minimum, stereotypes were not always inaccurate. That research is discussed next.

EARLY SOCIAL SCIENCE SCHOLARSHIP SUGGESTING THAT STEREOTYPES MAY NOT ALWAYS BE INACCURATE

Even Allport (1954/1979), who defined stereotypes as inaccurate (see Chapter 15), did *not* believe that *all* beliefs about groups were inaccurate. If we had “solid data” on how two or more groups differed, then a person who believed in those group differences might be accurate (which, for Allport, was the same as stating that they did not stereotype those groups). Allport, in contrast to much of the subsequent scholarship defining stereotypes as inaccurate (see Chapter 15), at least was logically coherent. For him, accurate beliefs about groups could exist but were not stereotypes.

The famous anthropologist Margaret Mead (1956) argued that people from different national or cultural backgrounds often possess general characteristics that distinguish them from other groups and that stereotypes partially, but incompletely, reflect these real differences. Similarly, sociologist Mackie (1973) argued that defining stereotypes as always inaccurate was inherently problematic and constituted little more than “arriving at truth by definition” (the title of Mackie’s article).

Much of the early (1940s to 1960s) evidence on stereotypes and intergroup perceptions showed that there was often widespread agreement among different groups about the characteristics of particular target groups. This included studies of Pakistani, Arab, American, Japanese, Chinese, Korean, Filipino, Samoan, African American, and Hawaiian groups (see Ottati & Lee, 1995, for a review). By the 1950s, social scientists began to recognize that these levels of agreement were high enough to begin considering the possibility that stereotypes really did sometimes reflect some degree of real differences between groups.

THE EXAGGERATION HYPOTHESIS

Variations on the idea that there might be some truth to stereotypes became known as the “earned reputation” theory and the “kernel of truth” hypothesis, both of which emphasized that, although stereotypes were largely inaccurate exaggerations, they did contain “a kernel of truth” (Allport, 1954/1979; Campbell, 1967; Tajfel, 1981; see McCauley, 1995, for a critical review of evidence on the exaggeration hypothesis). I do not know whether those promoting this idea thought about it in the following manner, but it always brought to my mind an image of a single kernel of decent corn (the “kernel of truth”) in an otherwise entirely rotten cob (the rest of the stereotype exaggerating and distorting that truth). Still, one kernel is better than none.

The exaggeration hypothesis has long and deep roots within social psychology. It long was the only perspective that permitted researchers to acknowledge that people were not always completely out of touch with social reality while simultaneously allowing researchers to position themselves well within the long-standing traditions emphasizing stereotype error and bias. Please keep this hypothesis in mind throughout Chapter 17 as it reviews in depth the findings of research on the accuracy of stereotypes.

THE FATE OF THE EARLY HINTS AND WHISPERS THAT STEREOTYPES ARE NOT ALWAYS INACCURATE

Nearly all of this early scholarship pointed in the same direction—that, frequently, many stereotypes had at least some degree of accuracy. Nonetheless, this research had no effect whatsoever on the conception of stereotypes as irrational reflections of bigotry in popular culture. Even more startling is that, until the 1990s, it had a nearly equally nonexistent effect on most social scientific views of stereotypes (see, e.g., the discussion of stereotypes that appears in almost any graduate or undergraduate psychology text from the 1980s and early 1990s and in prestigious and influential Handbook chapters, Annual Review chapters, and the like). As a graduate student in the 1980s, few of us were trained to carefully and evenhandedly evaluate the validity of people’s beliefs about groups. Instead, it was simply taken for granted that stereotypes were inherently inaccurate. And it was widely assumed that the reasonable and appropriate thing for a good social psychologist to do was study how and when those stereotype inaccuracies and biases manifested.

This is not just my opinion. Prominent psychological articles and texts from the 1970s, 1980s, and early 1990s are peppered with claims emphasizing the unjustified or inaccurate nature of stereotypes (e.g., Fiske & Taylor, 1984, 1991; Jones, 1986; Pinderhughes, 1989; Snyder, 1984; Snyder, Tanke, & Berscheid, 1977).² As Chapters 5 and 15 documented, they

are also quite common today. These claims were typically made without citation of articles demonstrating inaccuracy, which reflects the widespread agreement that one could take for granted as fact that stereotypes were inaccurate. Just as one need not cite evidence to support the claim that “the sky is blue,” one needed no research citations to support the claim that “stereotypes were inaccurate.”

The reasons for this are many and varied. Some are theoretical, others political and ideological (see Chapters 10 and 15). In addition, because so much of the evidence of accuracy came from outside psychology, many psychologists may have simply been unaware of it. Furthermore, even if they were aware, much of the evidence was indirect and suggestive, rather than clear and conclusive. The misdirection provided by Katz and Braly (1933)—of assuming that agreement among perceivers reflected something pernicious and evil about stereotypes, instead of likely reflecting accuracy—allowed for easy (mis)interpretation of the anthropological research showing high levels of agreement about groups. Little of the early evidence used objective measures against which to assess the accuracy of people’s stereotypes. Thus, all that was left was agreement, which could be dismissed as simply shared cultural myths about gender, ethnic, or national groups.

All that began to change in the late 1970s. Although only a very small number of psychological researchers were willing to tackle the thorny and controversial stereotype accuracy issue, they began to do so using a variety of measures: Census data, other objective measures, self-reports, standardized tests, and meta-analyses. Thus, this research was not as readily dismissed as was the earlier research. By the 1990s, this trickle of research began to become, if not quite a flood, at least a steady flowing stream.

Before reviewing that literature, however, it is necessary to lay down some foundations. First, so much research has addressed issues of stereotype bias and distortion that one might presume that there is already a vast literature addressing stereotype accuracy. Such a presumption is false, and the next sections explain why. Furthermore, as described in Chapters 10 through 12, assessing accuracy is generally a complex endeavor. The next sections, therefore, also explain how modern research generally assesses the accuracy of stereotypes. It also presents and justifies standards for characterizing any particular stereotype as accurate or inaccurate (something very rare within the entire social science literature on stereotypes).

These issues are important in their own right, at least for anyone interested in understanding how to assess the accuracy of any particular stereotype, and the principles laid down here should be applicable to many new situations beyond those specifically covered in this book. After all these foundations are laid down for understanding and studying the (in)accuracy of stereotypes, Chapter 17 reviews the empirical research that has assessed stereotype (in) accuracy.

CRITERIA FOR INCLUSION

To be included in Chapter 17, the empirical studies assessing the accuracy of stereotypes needed to meet two major criteria. First, they had to relate perceivers’ beliefs about some sort of target group with some sort of measure of what that group was actually like. This may seem obvious, but the social psychological discourse on stereotypes has often drawn conclusions about the inaccurate or unjustified nature of stereotypes based entirely on evidence addressing social cognitive processes—illusory correlations, priming, expectancy effects,

rationalizations of prejudice or inequalities, attributional patterns, etc. (see, e.g., the discussions of stereotyping in Aronson, 1999; Devine, 1995; Fiske, 1998; Fiske & Neuberg, 1990; Fiske & Taylor, 1984, 1991; Gilbert, 1995; Jones, 1986, 1990; Jost & Kruglanski, 2002; Nelson, 2002; and, indeed, almost any text or review of the social psychology of stereotypes that emphasizes their inaccuracy).

The view taken here, however, is that such research, although important on its merits, does not directly address accuracy, which can only be assessed by comparing belief to criteria (see Chapter 10 for a fuller elucidation of this rationale). Thus, to be included in Chapter 17, a study had to compare people's beliefs about one or more real groups composed of real people to some measure of the real characteristics of those real people. Studies assessing people's judgments regarding fictitious targets, which may be appropriate for testing hypotheses about basic judgmental processes (e.g., Brewer, Dull, & Lui, 1981; Hamilton & Rose, 1980), are not capable of addressing accuracy.

Second, studies needed to use an appropriate target group. Sometimes, researchers have, for example, asked people for beliefs about a group and used as criteria the characteristics of a haphazard sample of convenience (e.g., Allen, 1995; Dawes, Singer, & Lemons, 1972; Martin, 1987; Terracciano et al., 2005). These studies have an important disconnect between the stereotype they are assessing and the criteria they use.

This can, perhaps, be best illustrated with a concrete example. Let's say I assess the sex stereotypes held by students in one of my large lecture classes. I also plan to use their self-perceptions on the same characteristics as the criteria for real differences. I can then simply compare their stereotypes to the overall or mean self-perceptions in my class, can't I? Well, I can do the comparison, but it is not clear what that comparison will tell us. The men and women in an introductory psychology class may not be all that similar to a representative sample of men and women (those in my class are most likely, among other things, younger, healthier, thinner, less conscientious and more politically leftwing than would be found in a representative sample of American men and women). Thus, as perceivers, the students in my class may not look all that accurate—not because there is anything wrong with their stereotypes, but because their stereotype (of men and women in general) refers to a different group than I am using as my criteria (the men and women in my class). Furthermore, perhaps the young men and women taking a social science course were more similar to each other in their backgrounds, aspirations, and achievements than are the men and women in a nationally representative sample. If this were true, even if these perceivers accurately judged the real differences between men and women in general, such a study might *erroneously* conclude that they exaggerated real differences. This would occur because the criterion sample (the students in my class) actually has fewer sex differences than would a nationally representative sample. The bottom line is that criteria have to be appropriate to the stereotype, and, if it is not (with one exception, discussed later), it is not included in this review.³

Different Aspects of Stereotype (In)Accuracy

There are four broad ways in which the modern research has examined the accuracy of stereotypes. This breaks down into two types of accuracy (discrepancy from perfection or correspondence with differences) and two types of stereotypes (personal or consensual). Each of these is briefly discussed next.

DISCREPANCY FROM PERFECTION

Discrepancy from perfection refers to how close people come to exactly nailing on the head some level of some characteristic(s) in a group. It is assessed with discrepancy scores. For example, if Bernie says the average height of adult American women is 5 feet 5 inches, and it is really 5 feet 6 inches, Bernie underestimates their height by 1 inch. Chapter 12 discussed the use of discrepancy scores at some length, including Judd and Park's (1993) componential system for examining the accuracy of stereotypes. Discrepancy scores tell us how far off people are from perfect bull's eyes in estimating some characteristic(s) of a large group.

CORRESPONDENCE WITH DIFFERENCES

Correspondence with differences refers to how well people detect either variations between groups on some set of attributes or variations within groups on some set of attributes. For example, Sylvia might estimate the proportion of men and women with high school, college, and graduate degrees. Let's assume that she estimates that 1% more women than men receive a high school degree, 5% more women than men receive college degrees, and 10% more men than women receive graduate degrees. If those differences correspond to the real differences, then the correlation between her perceived and real differences will be very high, possibly even 1.0 (which could occur, e.g., if she had just perused recent Census data). Note, however, this also could occur even if she overestimates or underestimates both groups' likelihood of receiving degrees (see Chapter 12).

Similarly, one can examine beliefs about a single group by correlating beliefs about that group with the criteria. For example, one could correlate Yuan's beliefs about the likelihood of men graduating from high school, college, and graduate school with the actual likelihoods. This would not tell us anything about how accurately Yuan perceives sex differences, but it would tell us a lot about how accurately he perceives men's educational attainments.

PERSONAL STEREOTYPES

Personal stereotypes are the beliefs about groups held by a particular individual. Fred's view of Californians is uniquely Fred's; it may be identical to or completely different from anyone else's views. Personal stereotypes are assessed whenever individuals are asked to indicate their beliefs about groups. My examples above, in the sections describing discrepancy from perfection and correspondence with differences, all involved personal stereotypes because they all used as a hypothetical example one person's beliefs about men and women. One can, of course, assess different aspects of the accuracy of personal stereotypes by examining either discrepancies from perfection or correspondence with differences.

CONSENSUAL STEREOTYPES

In contrast, consensual stereotypes are the overall, or average, beliefs about a group held by some group of perceivers. For example, many people in a study may be asked to predict how much, per year, they believe that Jews, Catholics, Protestants, Muslims, and atheists donate to charity. Their average estimates for each group constitute a way to assess the consensual stereotype held by that group.

Consensual stereotypes have a uniquely important place in both social scientific and lay views of stereotypes, which frequently claim or assume that stereotypes are widely shared (e.g., Allport, 1954/1979; Claire & Fiske, 1998; Jost & Banaji, 1994; Katz & Braly, 1933; Marger, 1994; Pickering, 2001). Starting with Katz and Braly (1933), this idea has often been used in the service of the argument that stereotypes are little more than shared cultural myths. In this context, examining what the existing research shows about the accuracy of consensual stereotypes will be very interesting. The assessment of their accuracy, however, is nearly identical to that of personal stereotypes and can be accomplished both by using discrepancy scores (to assess deviation from perfection) and by assessing correspondence with differences. The only difference is that group means are used as perceptions, rather than the perceptions of a single individual.

What Is a Reasonable Standard for Characterizing a Stereotypic Belief as “Accurate”?

There is no objective gold standard with which to answer this question. So, the issue is, what is a reasonable standard? One hundred percent perfection is not reasonable and rarely occurs in any walk of life, including scientific theory. Of course, because there are two broad types of accuracy—discrepancy from perfection and correspondence with real differences—there needs to be two separate standards. Each is discussed next.

DISCREPANCIES

The bull’s eye. A good metaphor for accuracy comes from the best shot in target practice—commonly known as a bull’s eye. A bull’s eye is as good as it gets in target practice. In competition, a shot that is near the perimeter of the bull’s eye counts as much as one that is dead on center. Bull’s eyes are not tiny geometric points; they usually have width, which means one can hit a bull’s eye without being Robin Hood, who could hit the target dead center, then split his own arrow on the next shot.

So, my answer is that, for the type of social perceptual phenomena usually studied by social psychologists, a bull’s eye is within 10%. It is very conventional. In school in the United States, get 10% or less wrong, and one will typically receive an A—the highest grade possible. There are times when being 10% off could be terrible (e.g., estimating the difference between you and the stopped car in front of you when traveling at 65 mph), but those times are few and far between. If you are a professor who expects your new graduate recruit to complete the PhD program in 5 years and the person takes 5.5 years, you will probably not feel like you have been a failure as a mentor. If you estimate that 500,000 people of Pakistani descent live in the United Kingdom, and the real number is 545,000, or 450,000, you are likely to feel that you were just about right. And, frankly, even if you do not feel that way, I do characterize your estimate as pretty darn good.

There is nothing magic about 10%, and reasonable people may disagree. In *some* (rare) contexts, I would disagree myself. It is, nonetheless, the standard I will use in Chapter 17 to characterize a stereotyped belief as accurate. Especially when judging proportions and probabilities, as is common in the study of stereotype accuracy, within 10% is doing pretty well.

Some studies, however, do not report their results as percentages. Most that do not, however, do report their results as effect sizes or can be readily translated into effect sizes—real and perceived differences between groups in standard deviation units. Unfortunately, standard deviation units have no intuitive meaning to the statistically uninitiated. Nonetheless, they can be roughly translated into percentages. If Kay perceives Group A as .25 of a standard deviation (SD) higher on some attribute than Group B, this means that Kay perceives the average person in Group A to score higher on that variable than 60% of the people in Group B. Bingo! Ten percent difference. Therefore, for studies assessing stereotype accuracy using effect sizes, I characterize a perceived difference as accurate if it is within .25 SD of the real difference.

One last note. My standards often do not correspond to those used by the original authors (and you should read the original papers if you want to find out more about their standards). McCauley's research (see Tables 17-1 and 17-2) often used "less than 10%" off as his criterion for accuracy; we differ by a single percent, because I characterize 10% off as accurate. Others used statistical significance as their standard (e.g., if the perceived difference statistically exceeded or underestimated the real difference, they concluded it was not accurate). Although these standards have their own advantages and disadvantages, discussing them is beyond the scope of this chapter. My standards are simple and straightforward—within 10% (or .25 SD, which roughly translates into the same thing) and you are near enough for me to call you accurate.

Near misses. Accuracy is a matter of degree—it is not all or none (see Chapter 10 for a more detailed discussion). Therefore, it does not seem reasonable to characterize a belief that is 10% off as "accurate" and one that is 10.1% off as "inaccurate." So, how should we characterize near misses? As "near misses." A near miss is not accurate. But it is not too far off. Continuing with the archery metaphor, one can still rack up some points if one hits the target, even if one does not hit the bull's eye—not as many points as when one hits the bull's eye, but more than if one misses the target completely.

What, then, is a reasonable standard for a near miss? I will use the following: more than 10% off, but no more than 20% off. Within 20% is certainly not a bull's eye, but it is not completely out of touch with reality, either. It is certainly far more accurate, say, than being 40% off or more.

Again, reasonable people may disagree with characterizing more than 10% but no more than 20% off as near misses. Still, when trying to gain understanding of how accurate people's stereotypes are, knowing that their beliefs are near misses, rather than bull's eyes or completely inaccurate, seems pretty important.

Following the same rationale, then, as for accuracy, when results are only reported in standard deviations, I will use "more than .25 SD but no more than .50 SD" as my criterion for near misses. If Tom believes there is a .5 SD difference between groups, he believes that the mean of one group exceeds the scores of about 70% of the members of the other group. Again, a 20% difference.⁴

Types of discrepancies. The literature has focused on two broad types of discrepancies. By far, the most interesting and important discrepancy involves perceiving differences between groups. Do people perceive a larger or smaller difference between groups than really exists? Or do they perceive the difference accurately? These types of discrepancies directly test the exaggeration hypothesis that has been so long emphasized in the scholarly literature on stereotypes. It is also important for practical reasons. These discrepancies, when they show that

people exaggerate real differences on socially desirable attributes, indicate whether people unjustifiably perceive one group as “better” than another (more intelligent, more athletic, etc.). When they show that people underestimate real differences on socially desirable attributes, they indicate that people unjustifiably see groups as more similar to one another than they really are.

There is a second type of discrepancy reported in the literature that is still relevant as “inaccuracy” but has considerably less theoretical or practical importance. Independent of perceiving *differences*, sometimes people have a general tendency to overestimate or underestimate the *level* of some attribute(s). For example, let’s say men and women in the United States average 72 and 66 inches in height, respectively. Fred, however, believes that men and women average 74 and 68 inches, respectively. He consistently overestimates height by 2 inches (this is a fairly meaningless elevation effect—see Chapter 12), but he does not exaggerate sex differences in height.

CORRESPONDENCE WITH REAL DIFFERENCES: HIGH ACCURACY

How much correspondence should be considered “accurate”? Again, this is a judgment call. Nonetheless, I advocate holding people to a high standard—the same standards to which social scientists hold themselves.

Cohen (1988), in his classic statistical treatise imploring social scientists to examine the size of the effects they obtained in their studies and not just the “statistical significance” of the results, suggested that effect sizes above .8 could be considered “large.” Such an effect size roughly translates into a correlation of .4. By this standard, correlations of .4 and higher could be considered accurate because they represent a “large” correspondence between stereotype and reality.

This standard has been supported by two recent studies that have examined the typical effect sizes found in clinical and social psychological research. One recent review of over 300 meta-analyses—which themselves included over 25,000 studies and over 8 million human research participants—found that mean and median effect sizes in social psychological research were both about .2 (Richard, Bond, & Stokes-Zoota, 2003). Only 2.4% of social psychological effects exceeded .3. A similar pattern has been found for the phenomena studied by clinical psychologists (Hemphill, 2003). Psychological research rarely obtains effect sizes exceeding correlations of about .3. Accuracy levels (effect sizes) of .4 and higher, therefore, constitute a strong standard for accuracy.

As a general guideline, therefore, I will use a correlation of .4 between stereotype and reality as the cutoff for considering that stereotype accurate; .4 is double the typical effect size obtained in most social psychological studies, so it means that we are holding people to twice the standard to which we social psychologists hold ourselves. Last, according to Rosenthal’s (1985) binomial effect size display, a correlation of at least .4 roughly translates into people being right at least 70% of the time. This means they are right more than twice as often as they are wrong. That seems like an appropriate cutoff for considering the stereotype pretty accurate.

CORRESPONDENCE WITH REAL DIFFERENCES: MODERATE ACCURACY

Moderate correspondence, of course, is less than high correspondence. It reflects a mix of accuracy and inaccuracy. Following the same standards as science (Cohen, 1988; Richards

et al., 2003), I will characterize correlations between people's beliefs and reality ranging from .25 to .4 as moderately accurate. Such correlations do not reflect perfect accuracy, but nor do they reflect complete inaccuracy. Using Rosenthal's (1991) binomial effect size display, a correlation of .3, for example, means that people are right almost two-thirds of the time. Now, this also means they are wrong a little over one-third of the time. But two out of three ain't bad.

SOME MORE CAVEATS AND CLARIFICATIONS

The only aspects of the studies included in Chapter 17 that I discuss are those that involve the accuracy of stereotypes. Many of the studies addressed many issues other than accuracy. Those are all beyond the scope of Chapter 17, which is not intended to be a comprehensive review of all the information presented in all of the studies. It is only a review of their findings regarding the (in)accuracy of stereotypes.

None of the studies described in Chapter 17 use my exact 2×2 terminology of personal and consensual stereotypes, which can be evaluated using either discrepancies or correspondence with real differences (or both). Often, they simply discuss "stereotypes." Regardless, I do make that distinction and describe their results accordingly, regardless of whether they described their results this way.

Occasionally, the original authors do distinguish between personal and consensual stereotypes, although they generally use somewhat different terminology than I do. For example, consensual stereotypes are sometimes discussed as "aggregated" results or stereotypes (because they aggregate across all perceivers). Personal stereotypes are sometimes discussed as "individual" stereotypes; and the Judd/Park/Ryan group uses the term "within subject sensitivity correlations" to refer to what I call "personal stereotypes: correspondence with real differences." The main point here is that it is important to distinguish between consensual and personal stereotypes, and that accuracy can be assessed as either discrepancies from perfection or correspondence with real differences (which provide different, not better or worse, information about accuracy). Chapter 17 makes those distinctions, whether or not they appeared in the original articles.

Notes

1. This corresponds to what is frequently called the dyadic level of analysis (e.g., Jussim, 1991; Kenny, 1994), because there is one perceiver and one target. However, the discussion of all three levels of analysis here has assumed a single perceiver and that what is varying is the number and nature of the targets (population, small group, individual). Therefore, "person perception," is discussed here as occurring at the individual level of analysis.

2. Not all of these claims took the form of a blunt declaration that stereotypes were inaccurate. Some (e.g., Fiske & Taylor, 1984, 1991 provided an ostensibly neutral definition of stereotypes and then went on to discuss stereotypes in an almost entirely pejorative manner (see Chapter 15).

3. Amazingly, most of these studies also provided considerable evidence of accuracy. For example, the consensual stereotype accuracy correlations for the studies described in McCauley (1995) and Martin (1987) ranged from about .6 to over .9. Even in Allen (1995), which is titled "Gender Stereotypes Are Not Accurate . . .," those correlations are over .3.

4. Standard deviations, as the statistically inclined are well-aware, are not linear. Therefore, .52 SD comes closer to capturing a 20% difference than does .50. But .50, as a round number, is easier to use and remember, and ease of use has its own value. An SD difference of .50 actually means the mean of one group is higher than the mean of 69.15% of the members of the other group. Close enough for me.

*That all men are equal is a proposition which, at ordinary times,
no sane individual has ever given his assent.*

— ALDOUS HUXLEY

17 Pervasive Stereotype Accuracy

Warning: Turn Back Now, Before It Is Too Late

This chapter contains content that may be deeply upsetting to anyone committed to the view of stereotypes as inherently or generally inaccurate and irrational. If you have read this book continuously, you undoubtedly do not need these warnings and know what to expect. However, these warnings are necessary for anyone reading this chapter without reading the rest of the book.

Warning I: DO NOT READ THIS CHAPTER without having first read Chapters 10 through 12, 15, and 16. You will need those chapters to understand what I mean by accuracy generally and when I describe the results of the studies reviewed below as showing that people's beliefs were "accurate," "near misses," or "inaccurate" in this chapter.

Warning II: DO NOT READ THIS CHAPTER unless you are willing to consider the possibility that stereotypes are often accurate. DO NOT READ THIS CHAPTER if you think that merely considering the possibility that many of people's beliefs about groups (stereotypes) have a great deal of accuracy makes someone a racist, sexist, etc. DO NOT READ THIS CHAPTER if you believe that stereotypes are inherently inaccurate, flawed, irrational, rigid, etc., *and* that this belief cannot be or should not be revised if empirical scientific data fail to fully support it.

Introduction to the Review of Research on Stereotype Accuracy

The research on stereotype accuracy has several extraordinary features. As documented in Chapter 15, research in the social sciences has considered beliefs about almost any type of

group (demographic groups, occupational groups, political groups, memberships in organizations, etc.) to be a stereotype, and it has considered all sorts of beliefs (personality traits, achievements, behaviors, attitudes, etc.) to be parts of stereotypes. Consistent with this modern idea that any belief about any group is a stereotype, a major strength of the stereotype accuracy research is that it has examined all sorts of groups and beliefs about all sorts of attributes.

Of course, each study, individually, has important imperfections and limitations, and these are duly noted and discussed below. Thus, it is possible for intelligent readers motivated to deny stereotype accuracy to come up with “yes, buts” for each study (see Chapter 16 for a discussion of “yes, buts”).

And I would probably agree with most of them, at least in the narrow sense of applying them to a particular study. But as an attempt to dismiss the whole area of research, those “yes, buts” fail because of the extraordinary diversity of studies and their extraordinary similarity in results. Do some studies address stereotypes about groups into which people self-select (political parties, sororities, etc.)? Yes. Are those types of stereotypes more accurate than those regarding, say, sex and race? Well, the many studies of the accuracy of sex and race stereotypes are reviewed below—just read them and compare for yourself. Did some studies examine stereotype beliefs about which it is fairly easy to obtain clear objective information (such as sex distribution into various jobs)? Yes. Are those types of stereotypes more accurate than those regarding more fuzzy and difficult-to-observe attributes, such as personality and attitudes? Again, both types of studies have been performed and are reviewed below—see for yourself. Do studies using self-reports as criteria yield different results than studies using objective criteria, such as the Census? See for yourself.

Accuracy of Ethnic and Racial Stereotypes

Table 17–1 summarizes the results of all studies that have assessed the accuracy of racial stereotypes that I could find that met the criteria for inclusion described in Chapter 16. Each study is described next.

McCAULEY AND STITT (1978): THE FIRST STUDY OF THE ACCURACY OF RACIAL STEREOTYPES

McCauley and Stitt (1978) provided the first rigorous examination of the accuracy of people’s beliefs about differences between African Americans and other Americans. These beliefs included the percentage of African Americans and other Americans who were high school graduates, born illegitimately, unemployed last month, crime victims, on welfare, parents of four or more children, and in a household headed by a female. They examined these beliefs among five different samples: high school students, college undergraduates, master’s in social work (MSW) graduate students, members of a union, and members of a church choir (total $N = 62$). McCauley and Stitt (1978) did not examine the accuracy of personal stereotypes; all of their results focused on consensual stereotypes (see Chapter 16 for definitions of personal vs. consensual stereotypes).

In his review of research on the exaggeration hypothesis, McCauley (1995, Table 2) presented the data from his 1978 study in more detail. This is important because I was able to use this data to compute some results that were not reported in either the 1978 or 1995 papers.

McCauley and Stitt's (1978; McCauley, 1995) data were capable of addressing the accuracy of three different aspects of people's ethnic/racial stereotypes: (1) regarding Americans in general, (2) regarding African Americans, and (3) regarding the *differences* between African Americans and other Americans. Both discrepancies and correspondence could be addressed with their data and are discussed next (see Chapter 16 for definitions of different types of stereotype accuracy, including discrepancies from perfection and correspondence with real differences).

Consensual stereotype accuracy: Discrepancies. There were a total of 35 consensual discrepancies assessed (five groups by seven judgments) for each of the three types of stereotype accuracy. People's judgments of Americans were accurate (within 10%) 17 times, they had 13 near misses (within 20%), and they were inaccurate 5 times (more than 20% off). People's judgments of African Americans were accurate 17 times, they had 14 near misses, and they were inaccurate 4 times. Overall, there was an "elevation" (see Chapter 12) tendency—people tended to overestimate for both groups (i.e., estimate a higher percentage than indicated in the census).

So, there was some, but not much, evidence of inaccuracy here. But here is an even more surprising result: The consensual stereotype of *differences* between African Americans and other Americans was accurate 27 times and had 8 near misses. No judgments of differences were inaccurate. Furthermore, all eight of the near misses underestimated, rather than exaggerated, real differences between African Americans and other Americans.

Consensual stereotype accuracy: Correspondence with real differences. Their results regarding the extent to which beliefs about the groups, and group differences, corresponded with reality were even more striking. Based on the data in McCauley's (1995) Table 2, it is possible to compute correlations between the consensual stereotypes and Census data for each of the five samples. The correlation of people's stereotypes of African Americans with Census data ranged from .27 to .83 and averaged .60.¹ The correlation of people's stereotypes of Americans with Census data ranged from .84 to .97 and averaged .93. The correlation of people's perceptions of differences with real differences ranged from .87 to .90 and averaged .88. These are stunning levels of correspondence (correlational) accuracy.

Conclusion. Overall, therefore, this study provided clear evidence for the accuracy of some consensual racial stereotypes. It provided no support for the exaggeration hypothesis.

A major strength of the study was that, in contrast to the overwhelming majority of research on stereotypes, including many of the remaining studies cited in this chapter as well as much of the error and bias process work (e.g., Fiske & Neuberg, 1990; Jones, 1990), it included several noncollege student samples. Although not nationally representative samples, the stereotypes held by the different groups were highly similar, which bodes well for the likely generalizability of their findings.

Of course, the study also had important limitations. It did not assess the accuracy of individual stereotypes. It only assessed stereotypic beliefs that could be compared to Census data. It did not examine stereotypes other than those regarding race. So, let's see what some of the subsequent research showed.

TABLE 17-1

The Accuracy of Racial and Ethnic Stereotypes					
Study and Stereotype	Perceivers	Criterion	Predominant Pattern of Discrepancies ^a	Individual Correlations (Personal Stereotype Accuracy)	Aggregate Correlations (Consensual Stereotype Accuracy)
McCauley & Stitt (1978): beliefs about demographic differences between African Americans and other Americans	Five haphazard samples (church choir, union members, students, etc.), total <i>N</i> = 62	U.S. Census data	Accuracy	Not available	<i>Beliefs about</i> ^{b,c} : African Americans:.60 Americans:.93 Differences between African Americans and other Americans:.88
Ryan (1996): beliefs about differences in the personal characteristics of African American and White University of Colorado students	Random samples of 50 African American and 50 White University of Colorado students	Self-reports of the random samples of perceivers	Among Whites, accuracy; among African Americans, accuracy and exaggeration (tied)	African American perceivers:.42 ^b White perceivers:.36 ^b	African American perceivers: .73, .53, .77 ^{c,d} White perceivers: .77, .68, .72 ^{c,d}
Ashton and Esses (1999): beliefs about the achievement of nine Canadian ethnic groups	94 University of Western Ontario students	Board of Education achievement data	Accuracy: 36 of 94 Exaggeration: 33 of 94 Underestimation: 25 of 94 ^c	.69	Not available

Wolsko et al. (2000): beliefs about differences between African Americans and White Americans	83 White University of Colorado undergraduates	Objective data from government (e.g., Census) and other (e.g., National Basketball Association) sources	Underestimation	Not available	Not available
---	--	---	-----------------	---------------	---------------

^a Except where otherwise stated, all discrepancy results occur at the consensual level. Accuracy means within 10% of the real percentage or within 0.25 of a standard deviation. Exaggeration means that the perceived differences between groups exceeded the group differences on the criteria. Underestimation means that the perceived differences between groups was smaller than the group differences on the criteria. Except where otherwise noted, only one word is entered in this column when one pattern (e.g., “accuracy”) occurred for a majority of results reported. When there was no majority, the top two results, in order of frequency (most frequent first), are reported here.

^b For simplicity, if the study reported more than one individual level (average) correlation, I have simply averaged all their correlations together to give an overall sense of the degree of accuracy.

^c These correlations do not appear in the original article, but are computable from data that were reported.

^d For each group of perceivers, the first correlation is the correspondence between their judgments and the self-reports of their own groups; the second correlation is the correspondence between their judgments and the self-reports of the other group; and the third correlation is the correspondence between the perceived difference between the groups and the difference in the self-reports of the two groups.

^e These are personal discrepancies. Ashton and Esses (1999) computed a personal discrepancy score for each perceiver, and then reported the number of perceivers who were within 0.2 standard deviations of the criteria and the number that exaggerated real differences (saw a difference greater than 0.2 SD larger than the real difference) or underestimated real differences (saw a difference more than 0.2 SD smaller than the real difference).

McCauley and Stitt (1978), Ryan (1996), and Wolsko et al. (2000) examined beliefs about African Americans and White Americans. Wolsko et al. (2000) found statistically significant evidence of underestimation, but their data were not reported in such a manner as to be able to determine whether discrepancies were within 10% of the real percentage, or within 0.2 of a standard deviation. Ashton and Esses (1999) examined beliefs about nine different Canadian ethnic groups, and discrepancy results refer to the number of participants showing each pattern. The results reported here refer to their Table 2, which reports the number of perceivers within 0.2 of a standard deviation of the real difference. They did not report results from which the number of perceivers within 0.25 of a standard deviation of the real difference could be identified.

Ryan’s (1996) results refer to her stereotypicality results, not her dispersion results. Exaggeration means that the perceived differences between groups exceeded the group differences on the criteria. Underestimation means that the perceived differences between groups was smaller than the group differences on the criteria. See text for explanation of the color-blind and multicultural conditions of the Wolsko et al. (2000) study.

Individual correlations involve computing, for each individual perceiver, the correlation between their judgments (stereotypes) and the criterion. Studies performing this analysis typically report the average of those correlations. Aggregate correlations refer to the correlation between the overall average perceived difference between the groups (for the whole sample) and the group difference on the criteria.

RYAN (1996): THE ACCURACY OF AFRICAN AMERICAN AND WHITE STUDENTS' PERCEPTIONS OF ONE ANOTHER

Ryan (1996) examined racial stereotypes among University of Colorado students. First, she identified 17 attributes that were positive or negative and stereotypic and counterstereotypic for Whites and African Americans (attributes were a mix of behaviors, achievement, and personality and included athletic, likely to drop out of college, parental income and education, intelligence, and self-centered).

Then she selected random samples of African American and White University of Colorado students. Those students then (1) rated themselves on these 17 attributes and (2) rated African American and White University of Colorado students on those 17 attributes. These attributes included behaviors, personality, and achievements, such as athletic (positive, stereotypical of African American students and counterstereotypic of White students), sexually aggressive (negative, stereotypical of African American students and counterstereotypic of White students), academically intelligent (positive, counterstereotypic of African American students and stereotypic of White students), and self-centered (negative, counterstereotypic of African American students and stereotypic of White students). Ryan computed both discrepancy scores and correlations to assess the accuracy of the racial stereotypes held by her samples of African American and White students.

Consensual stereotype accuracy: Discrepancies. Although Ryan (1996) did not report patterns of discrepancies for individuals, she did report mean discrepancies for each rating, separately, by perceiver group and target group. These, therefore, are consensual discrepancies.

Ryan (1996) found differing patterns of discrepancies among her African American and White samples. Ryan used the Judd and Park (1993) method (summarized in Chapter 12), which involved separating out effects for perceiver group, target group, attribute stereotypicality, and attribute valence, to assess the accuracy of the average perceived differences. Although Ryan (1996) did not report her results in this manner, her Table 1 data can answer several questions about the consensual stereotypes held by African American and White University of Colorado students about one another. Using my system (see Chapter 16) of classifying stereotype beliefs as accurate, near misses, or inaccurate, her results were as follows:

1. African Americans' beliefs about African Americans were accurate five times, they had five near misses, and they were inaccurate seven times.
2. African Americans' beliefs about White Americans were accurate 3 times, they had 2 near misses, and they were inaccurate 12 times.
3. African Americans' beliefs about differences were accurate seven times, they had two near misses, and they were inaccurate eight times. Including the near misses, most of their inaccuracies exaggerated real differences. Seven of 10 inaccuracies exaggerated real differences; one underestimated real differences; once they perceived a difference where none existed (the real difference was less than 10%), and once they saw a difference in the wrong direction (seeing African Americans as higher on athleticism when Whites' self-reported athleticism exceeded that of African Americans' self-reported athleticism).
4. White Americans' beliefs about African Americans were accurate five times, they had five near misses, and they were inaccurate seven times.

5. White Americans' beliefs about White Americans were accurate 5 times, they had 2 near misses, and they were inaccurate 10 times.
6. White Americans' beliefs about differences were accurate nine times, they had four near misses, and they were inaccurate four times. Their inaccuracies showed no clear pattern. Including the near misses, they exaggerated real differences twice, they underestimated real differences three times, twice they saw differences when none existed, and once they saw a difference in the wrong direction (athleticism again).

Overall, therefore, these results show that:

1. The consensual stereotypes held by both African Americans and Whites had an intermediate degree of (in)accuracy with respect to judging the absolute levels of the characteristics of both groups. Although there were a fair number of bull's eyes, there were also many near misses, and even more inaccurate consensual discrepancies.

Inaccuracies occurred mainly because of a widespread tendency among perceivers in both groups to overestimate the levels of the various characteristics in both target groups. This explains why there were more inaccuracies in perceptions of each group than in perceptions of differences between the groups (if one overestimates both groups by the same amount, the difference will be right on target). This form of inaccuracy does not fit into any prior theoretical analysis of stereotype inaccuracy, and its source and meaning is unclear. It is probably a relatively meaningless "elevation" effect (see Chapter 12).

2. African Americans' consensual stereotypes regarding racial differences generally exaggerated real differences.
3. White Americans' consensual stereotypes regarding racial differences were generally accurate and showed no clear tendency to exaggerate real differences.

Consensual stereotype accuracy: Correspondence with real differences. Ryan (1996) reported data (in her Table 1) from which can be computed six separate consensual stereotype accuracy correlations (of beliefs with criteria). Three are for African American perceivers, and all three show high consensual stereotype accuracy: their beliefs about African Americans (.73), about White Americans (.53), and about the differences between African Americans and White Americans (.77). Three are for White American perceivers: their beliefs about African Americans (.68), their beliefs about White Americans (.77), and their beliefs about the differences between African American and White Americans (.72).

In addition to the strikingly high levels of accuracy in all of these correlations, there is one other notable pattern. Both African Americans and White Americans were somewhat better at judging their own ethnic/racial group (correlations of .73 and .77, respectively) than at judging the other ethnic/racial group (correlations of .53 and .68, respectively). Apparently, although these correlations show high consensual stereotype accuracy across the board, people may know their own group somewhat better than they know the other group.

Personal stereotype accuracy: Correspondence with real differences. Ryan (1996) also calculated the accuracy of *each* individual perceiver's beliefs, by correlating the ratings of each group with the group's self-reports. Her Table 4 (p. 1122) reports the average of these

correlations, by perceiver group, target group, stereotypic characteristics, and counterstereotypic characteristics. For African Americans, these averaged correlations ranged from .22 to .60 and averaged .42. For Whites, these averaged correlations ranged from .12 to .56 and averaged .36. These results indicate that most people were moderately accurate in their perceptions of how African Americans and Whites varied on these attributes.

Her results also showed that people were generally more accurate judging Whites (correlations ranging from .36 to .60) than African Americans (correlations ranging from .12 to .48). This is consistent with the idea that stereotypes arise from experience with reality. There are far more Whites than African Americans at the University of Colorado and, indeed, throughout the United States. It seems likely that many African Americans have more contact and experience with Whites than Whites have, on average, with African Americans (although this contact difference could reflect prejudice to some degree, this would be true even in the utter absence of prejudice or segregation simply on the basis of the far higher proportion of Whites). If most people have lots of experience with Whites and only some people have experience with African Americans, and if stereotypes are well-grounded in experience (when it is available), then stereotypes about Whites, in general, would be more accurate than stereotypes about African Americans.²

People were also more accurate judging counterstereotypic attributes (correlations ranging from .34 to 0.60) than stereotypic ones (correlations ranging from .12 to .40). Although it is not clear why these patterns occurred, one possibility is that this may represent political correctness seeping into people's reported beliefs. To the extent that people are aware of general stereotypes and see them as something bad, they may be reluctant to acknowledge seeing groups in stereotype-consistent ways. In contrast, there is little or no political correctness pressure to see groups as counterstereotypic.

Conclusion. Although Ryan's (1996) research can be criticized for using self-reports (see also Chapter 11), in conjunction with McCauley and Stitt's (1978) study, I consider it a strength. As discussed in Chapter 11, if it looks like a duck, walks like a duck, and sounds like a duck . . . Now we have two different studies, conducted decades apart, asking about very different sets of beliefs, using very different criteria, and yielding *highly similar* results regarding consensual stereotypes. The consensual stereotype corresponded very highly with the criteria (Census data in one study, self-reports in another).

This was also the first study to examine the accuracy of personal racial stereotypes. The correlations of .36 and .42 indicated that, on average, people's personal stereotypes corresponded moderately well with target group members' self-reports.

A limitation of the study is that Ryan (1996) did not examine the discrepancy (in)accuracy of the personal stereotypes held by her individual participants. Therefore, although we know how discrepant the consensual stereotypes were from the criteria, we cannot determine how discrepant the individual perceivers' personal stereotypes were from the criteria.

Her consensual discrepancy analyses, however, found that White perceivers' beliefs about racial differences were mostly accurate, without much of a clear tendency to over- or underestimate the real differences between the groups. Although her African American sample was also reasonably accurate in perceiving differences by my 10% standard, she did find that they also tended to systematically exaggerate the real differences between the groups.

This result is very interesting. In Ryan's (1996) study, it was African Americans', rather than Whites', stereotypes that were most biased. This pattern is not consistent with perspectives

suggesting that stereotype biases function largely to support or justify the power or position of higher status groups (e.g., Jost & Banaji, 1994; Sidanius & Pratto, 1999). Instead, Ryan (1996) argued that this pattern occurred because groups are most likely to exaggerate real differences when their group identity is threatened in some way (and the historical oppression of African Americans constitutes just such a threat). Thus, perhaps exaggeration is not a general or defining characteristic of stereotypes, and, perhaps, it is not even a common characteristic of the stereotypes held by high-power, high-status groups. Instead, exaggeration may be more likely to appear when people's group identities are threatened (see also Fein & Spencer, 1997).

ASHTON AND ESSES (1999): STEREOTYPES ABOUT ETHNIC DIFFERENCES IN ACADEMIC ACHIEVEMENT

Ashton and Esses (1999) examined beliefs about differences in academic achievement among nine different Canadian ethnic groups: Native Indians, British, Canadian-born Black, Caribbean-born Black, Chinese, East Indian/Pakistani, Jewish, Portuguese, and Vietnamese. Ninety-four University of Western Ontario undergraduates estimated the average achievement for members of these groups attending high school in Toronto. They estimated the high school grades for these groups, using the same grading scale that is used throughout Canadian high schools (and one condition for involvement in the study was that the participant had to graduate from a Canadian high school). Thus, the stereotype measure was quantitative, relatively objective (at least when compared to, e.g., the type of self-reports used as criteria in Ryan's [1996] study), and on a scale highly familiar to all participants. Furthermore, the criteria were similarly quantitative and objective—reports including the average grades published by the Toronto Board of Education.

Consensual stereotype accuracy: Discrepancies. Ashton and Esses' (1999) Table 4 reported consensual discrepancies—the average difference between estimated and mean achievement for each of the nine target groups. These discrepancies indicated near bull's eye levels of accuracy for most groups. The biggest discrepancy was the underestimation of Jews' achievement by 3.6 points (on a scale going from 0 to 100, in which 60% of the students averaged between 60 and 80). This discrepancy was 0.3 SD off, so, although it would not be characterized as accurate using my 0.25 SD cutoff, even this highest discrepancy is a near miss (0.5 SD cutoff) rather than completely inaccurate. All other discrepancies were bull's eyes. Unfortunately, Ashton and Esses (1999) neither assessed the extent to which the consensual stereotypes corresponded with real differences nor reported the type of data from which such correspondence could be assessed.

Personal stereotype accuracy: Discrepancies. This is one of very few studies to report results regarding the discrepancies from perfection for individual participants. They did not, however, assess the accuracy of the perceivers' achievement stereotypes for each group; nor did they report results from which this information could be discerned. Instead, they assessed the accuracy of perceivers' judgments of the variability in achievement across the nine ethnic groups. This answers the question: Do people systematically exaggerate real differences between groups? If they do, then people's stereotypes should be more variable (more extreme) than the real differences. The standard deviation is a common statistical measure of variability, so, for each perceiver, they simply compared the standard deviation of the stereotypes of the nine groups to the standard deviation of the real differences between groups.

Because there is no single standard for what constitutes accuracy, they used five different criteria: within 0.2, 0.4, 0.6, 0.8, and 1 standard deviation. Note that none of these quite corresponds to my definition of accuracy, which is within 0.25 of a standard deviation. So, let's first look at their results for people within 0.2 and 0.4 of a standard deviation.

Using the 0.2 criteria, 36 people were accurate, 25 underestimated real differences, and 33 exaggerated real differences. Using the 0.4 criteria, 52 people were accurate, 17 underestimated real differences, and 25 exaggerated real differences. We can estimate their results for a 0.25 cutoff by interpolating (moving one fourth of the way from 0.2 to 0.4). Doing so yields the following approximations (and please keep in mind that these are just approximations) for my standards for accuracy: 40 people were accurate, 23 underestimated real differences, and 31 exaggerated real differences. By any standard, these results show that at least a substantial plurality of perceivers' judgments were quite accurate, and there was only a slightly greater tendency to exaggerate than underestimate real differences. Even this pattern disconfirms exaggeration as a *defining* or *essential* characteristic of stereotypes (see Chapter 15).

Ashton and Esses (1999) also performed a set of analyses that help shed some light on who was most and least likely to be accurate. In addition to the stereotype questions, perceivers also completed the Right Wing Authoritarianism (RWA) scale and two intelligence tests. For these analyses, they considered anyone who was within 0.6 SD of the real differences to be accurate. The 14 people who underestimated real differences (the standard deviation of their stereotypes was 0.6 SD less than the standard deviation of the real differences) scored very low on the RWA scale. Because the RWA scale is highly correlated with liberal versus conservative ideology (higher RWA, more conservative; Altemeyer, 1981), those scoring very low on the RWA are likely to be extremely liberal in their politics. One interpretation of this finding, therefore, is that those most likely to inaccurately underestimate real differences were liberals in denial about real group differences. Another aspect of their results makes this pattern even more amazing—intelligence did not matter for this group. Brainy liberals were just as likely as dumb liberals to inaccurately minimize real differences. This pattern suggests a link between liberal politics, a denial of difference ideology (discussed in Chapter 10; see also Ryan, 2002; Wolsko, Park, Judd, & Wittenbrink, 2000), and inaccuracy in social stereotypes.

The 80 other participants scored much higher on the RWA scale than did the underestimators. Whether they all actually qualified as conservatives is unclear, but it seems unlikely given that this was a college sample attending a major urban Canadian university. Nonetheless, they were not as liberal as the underestimators. Among this group, intelligence mattered. Those who exaggerated real differences were not very smart. The group that was accurate, however, had intelligence scores at about the sample average (remember, these were college students and so, on average, probably had higher intelligence than the general population) and RWA scores that were much higher than those of the underestimators.

So, as long as you were not an extreme liberal, intelligence mattered—smarter people had more accurate stereotypes. This, too, is broadly consistent with the idea that stereotypes are largely anchored in reality—in general, intelligent people are likely to have more knowledge and understanding of the world around them than are less intelligent people. Whether extreme liberals have this knowledge but are unwilling to admit to it or genuinely do not realize that groups often do really differ is unknowable from this study—and an interesting question for future research.

Personal stereotype accuracy: Correspondence with real differences. For each perceiver, Ashton and Esses (1999) then examined the accuracy of the perceived rank ordering of the achievement of the nine groups by correlating the estimated rank with the actual rank (after converting the means to ranks). The average correlation between the estimated and real rank ordering of the groups was .69. People knew quite a lot about which groups had higher and lower levels of achievement.

Conclusion. A major strength of the study was that it is the only one of which I am aware to rigorously assess the accuracy of the stereotypes regarding more than two ethnic groups. Nearly all of the strong, modern research on ethnic stereotypes focuses on those regarding White Americans and African Americans. By studying stereotypes regarding nine different Canadian ethnic groups among a Canadian sample, the study significantly strengthens our ability to reach conclusions about stereotypes in general.

Nonetheless, the study also has important limitations. First, it only examined a single attribute—high school achievement. Whether people would be equally accurate about other characteristics was not assessed. And, of course, their perceivers were all college students.

Despite these limitations, this study provided some of the strongest, clearest evidence of stereotype accuracy to date. The consensual stereotype discrepancies were minimal; most were bull's eyes. The personal stereotype discrepancies were accurate more often than they either underestimated or exaggerated real differences. There was a slightly greater tendency to exaggerate than underestimate real differences; the exaggerators were unintelligent and not liberal, and the underestimators were extremely liberal. Last, the accuracy of people's individual stereotypes at capturing the real rank order of the groups' achievement was extraordinary.

WOLSKO ET AL. (2000): COMPARING THE ACCURACY OF RACIAL
STEREOTYPES PRODUCED BY ADOPTING A COLOR-BLIND VERSUS
MULTICULTURAL MINDSET

Wolsko et al. (2000, Experiment 2) examined the accuracy of 83 White University of Colorado undergraduates' racial stereotypes by comparing their beliefs to objective data obtained from a variety of U.S. government and other sources. These included 16 attributes related to classic racial stereotypes regarding work ethic/laziness, intelligence, religiosity, criminality, etc., which were assessed by comparing beliefs about percent unemployed, attendance at religious services, SAT scores, percent arrested for tax fraud, etc., among the two racial groups. They used the Judd and Park (1993) system for analyzing discrepancy scores (summarized in Chapter 12).

Wolsko et al. (2000), however, also added a very unique and creative twist: Their instructions encouraged participants to adopt either a *color-blind* or *multicultural* perspective. In the color-blind condition, instructions indicated that "... intergroup harmony can be achieved if we recognize that at our core we are all the same, that all men and women are created equal, and that we are first and foremost a nation of individuals" (p. 638). In the multicultural condition, the instructions indicated that "... intergroup harmony can be achieved if we better appreciate our diversity and recognize and accept each group's positive and negative qualities." In essence, the color-blind instructions encouraged people to deny or ignore group differences; the multicultural perspective encouraged people to recognize group differences.

Consensual stereotype accuracy: Discrepancies. Wolsko et al. (2000) did not report discrepancy scores for each judgment, because in the Judd and Park (1993) system one adds or averages all the discrepancies across judgments to obtain overall patterns of elevation, accuracy, and bias. This means that they did not provide information regarding discrepancies for the 16 judgments separately, so that, unfortunately, I cannot provide you with a tally of bull's eyes, near misses, and inaccuracies.

Their Table 4, however, did report overall results (averaging over all judgments), and this showed that, overall, people did fairly well. The overall average discrepancies were bull's eyes when judging stereotypic attributes (5.4% overestimation in the color-blind group, 7.3% overestimation in the multicultural group) and near misses when judging counterstereotypic attributes (12.5% overestimation in the color-blind group, 10.9% overestimation in the multicultural group). Of course, because these are overall results averaged across the 16 items, they might, but do not necessarily, mean that people always had near misses or bull's eyes (overestimating one attribute by 30% and underestimating another by 30% means that, on average, there is no discrepancy). In addition, like Ryan (1996), their results also showed that perceivers overestimated *both* stereotypic and counterstereotypic characteristics for both African American and White target groups (another of those relatively meaningless elevation effects).

Their results also showed that, for both target groups, *perceivers overestimated counterstereotypic characteristics more than they overestimated stereotypic characteristics*. Stereotypicality scores minus counterstereotypicality scores is an index of "stereotypicality" in the Judd and Park (1993; see also Chapter 12) system. When counterstereotypic traits are overestimated more than stereotypic traits, it means that stereotypicality is "negative"—people see the group as *less* stereotypic than it really is. In other words, they *underestimated* how stereotypical both groups were. This form of inaccuracy is directly opposite of the exaggeration hypothesis.

Color-blind versus multicultural mindset: Does denying or emphasizing differences lead to more accuracy? Which mindset, then, led people to adopt the most accurate stereotypes? The answer is the multicultural/recognize group differences mindset, hands down. Both groups underestimated stereotypicality and real differences—but the amount of underestimation was reduced by nearly two-thirds among the multicultural group. This study, therefore, provided the clearest direct evidence to date that when people have a mindset emphasizing denial of differences, their beliefs about groups become *less* accurate than when people have a mindset emphasizing the importance of recognizing real differences.

Conclusion. In some ways, their main result showing more accuracy among those adopting the multicultural mindset is so obvious that some might attempt to dismiss it as trivial. But I do not see their result as trivial at all. It constitutes the first empirical demonstration that people willing to admit that groups differ can be more accurate than those unwilling to admit that groups differ. When well-meaning people take the philosophical/political/spiritual idea that "deep down, we are all the same" too literally and conclude that, therefore, the experiences, practices, behaviors, and personal attributes of groups do not differ much, it undermines the accuracy of those people's beliefs about groups. It is true that we all bleed and all humans share over 99% of their DNA. But that does not mean that the behaviors, beliefs, histories, cultures, attitudes, practices, or accomplishments of all groups are the same. Denial is not usually a good strategy for staying in touch with reality.

The major limitations of this study involve what it did not do. It did not report results separately for the 16 judgments. It provided no evidence regarding the accuracy of personal stereotype discrepancy scores, and it reported neither personal nor consensual correspondence with criteria correlations (nor did it provide the type of data that would permit someone like me from computing them). And it only examined racial stereotypes regarding two American groups.

The Accuracy of Gender Stereotypes

Table 17–2 summarizes the results of all studies that have assessed the accuracy of gender stereotypes that I could find that met the criteria for inclusion described in Chapter 16. Each study is described next.

SWIM (1994): ASSESSING THE ACCURACY OF CONSENSUAL GENDER STEREOTYPES WITH OBJECTIVE CRITERIA

Swim (1994) performed one of the first and clearest examinations of the accuracy of consensual gender stereotypes. She (1) assessed a total of 293 college students' beliefs about the size of sex differences on 17 (Study One) or 15 (Study Two) attributes (aggressiveness, helpfulness, SAT scores, etc.), (2) located every meta-analysis¹ assessing the difference between men and women on these attributes, and then (3) compared the students' gender beliefs to the meta-analyses.

Meta-analyses typically report differences in terms of the number of standard deviations of difference between two groups (in this case, males and females). Swim, therefore, translated perceivers' stereotypes into perceived standard deviations of difference and then compared those perceived differences to the meta-analyses. Swim (1994) did not examine the accuracy of individual stereotypes; instead, all of her results addressed the accuracy of the consensual stereotypes.

Consensual stereotype accuracy: Discrepancies. The consensual stereotypes were highly accurate. In Study One, 8 of 17 perceived differences were accurate; there were 4 near misses (3 of which underestimated real differences, and 1 of which was a reversal [people believed women had slightly higher verbal SAT scores when, in fact, men had slightly higher scores]); and 5 exaggerated real differences. In Study Two, 10 of 15 perceived differences were accurate; there were 3 near misses (all of which underestimated real differences) and 2 inaccuracies (both of which underestimated real differences). Across the two studies, there was a slightly greater tendency to underestimate than exaggerate real differences between males and females.

Consensual stereotype accuracy: Correspondence with differences. Swim (1994) also correlated the average perceived differences with the average real differences, as indicated by meta-analyses. In Study One, the consensual stereotypes correlated .79 with the real differences; in Study Two, the consensual stereotypes correlated .78 with the real differences. These are extraordinarily high levels of accuracy.

Conclusion. The use of meta-analysis as the criterion for accuracy constitutes a major strength of this study. Meta-analyses succinctly summarize all that is known to science about

TABLE 17-2

The Accuracy of Gender Stereotypes					
Study and Stereotype	Perceivers	Criterion	Predominant Pattern of Discrepancies ^a	Individual Correlations (Personal Stereotype Accuracy)	Aggregate Correlations (Consensual Stereotype Accuracy)
McCauley et al. (1988); McCauley and Thangavelu (1991): beliefs about the sex distribution into different occupations	College students, high school students, rail commuters (<i>N</i> = 521 over the 5 studies)	Census data on proportion of women employed in various occupations	Accuracy	Not available	.94-.98 ^b across five studies
Swim (1994) ^c : beliefs about sex differences on 17 characteristics	Introductory psychology students (<i>N</i> = 293 over two studies)	Meta-analyses of sex differences on 17 characteristics	Accuracy	Not available	Study One: .78 Study Two: .79
Briton & Hall (1995): beliefs about sex differences in nonverbal behavior	441 introductory psychology students	Meta-analysis of nonverbal sex differences	Accuracy	Not available	Female perceivers: .74 Male perceivers: .68

Cejka and Eagly (1999): beliefs about the sex distribution into different occupations	189 introductory psychology students	Census data on proportion of women employed in 80 occupations	Accuracy and underestimation	Not available	.91
Beyer (1999) ^d : beliefs about the sex distribution into different majors and mean GPA of men and women in those majors	265 college students	College data on proportion of men and women in different majors, and their GPAs Meta-analyses of sex differences on 77 characteristics	Accuracy and underestimation	<i>Proportion:</i> Male perceivers: .48 Female perceivers: .52 <i>GPA</i> Male targets: .22 Female targets: -.04 .43	<i>Proportion:</i> Male perceivers: .80 Female perceivers: .79 <i>GPA</i> Male perceivers: .35 Female perceivers: .34
Hall and Carter (1999): beliefs about sex differences on 77 characteristics	708 introductory psychology students		Not available		.79

(Continued)

TABLE 17-2

The Accuracy of Gender Stereotypes (Continued)					
Study and Stereotype	Perceivers	Criterion	Predominant Pattern of Discrepancies ^a	Individual Correlations (Personal Stereotype Accuracy)	Aggregate Correlations (Consensual Stereotype Accuracy)
Diekman et al. (2002): beliefs about the attitudes of men and women	617 college students over three studies	Attitude positions endorsed by men and women on the General Social Survey (random sample of American adults)	Accuracy for consensual discrepancies; near miss for personal discrepancies	Male targets: .45 ^c Female targets: .54 ^c When judging sex differences: .60	Male targets: .66 ^c Female targets: .77 ^c When judging sex differences: .80

^a Except where otherwise stated, all discrepancy results occur at the consensual level. Accuracy means within 10% of the real percentage or within 0.25 of a standard deviation. Exaggeration means that the perceived differences between groups exceeded the group differences on the criteria. Underestimation means that the perceived differences between groups was smaller than the group differences on the criteria. "Near miss" means perceivers were more than 10% wrong, but no more than 20% wrong.

Only one word is entered in this column when one pattern (e.g., "accuracy") occurred for a majority of results reported. When there was no majority (or the majority could not be determined -from their data), the top two results, in order of frequency (most frequent first) are reported here.

^b These correlations do not appear in the original article, but are computable from data that was reported.

^c Swim (1994) sometimes reported more than one meta-analysis as a criterion for a perceived difference. In that case, I simply averaged together the real differences indicated by the meta-analyses in order to have a single criterion against which to evaluate the accuracy of the perceived difference.

^d For Beyer (1999), all results are reported separately for men and women perceivers, except the individual correlations for GPA. Because there was no significant sex of perceiver difference in these correlations, Beyer reported the results separately for male and female targets.

^e For simplicity, if the study reported more than two correlations, I have simply averaged all their correlations together to give an overall sense of the degree of accuracy.

Individual correlations involve computing, for each individual perceiver, the correlation between their judgments (stereotypes) and the criterion. Studies performing this analysis typically report the average of those correlations. Aggregate correlations refer to the correlation between the overall average perceived difference between the groups (for the whole sample) and the group difference on the criteria.

some effect, or, in this case, sex differences. Swim (1994) found that perceived differences between men and women were typically of about the same magnitude as the real differences (the consensual stereotypes were not very discrepant from the real differences). Furthermore, the consensual stereotypes closely corresponded with the real differences—the more males and females actually differed, the more the males and females were seen to differ in the consensual stereotypes.

This study did, however, have two significant limitations. First, Swim (1994) provided no information regarding personal stereotypes (discrepancies or correspondence). Second, both samples were college students, so whether her obtained pattern of very high consensual accuracy holds in the general population is unknowable from her data.

BRITON AND HALL (1995): GENDER STEREOTYPES REGARDING NONVERBAL BEHAVIOR

Briton and Hall (1995) examined the accuracy of 441 college students' stereotypes regarding sex differences in 17 aspects of nonverbal communication (amount of talking, speech disfluencies, interruptions, smiles, etc.). They argued that these beliefs are important because nonverbal behavior is a significant aspect of interpersonal communication. Like Swim (1994), Briton and Hall (1995) used prior meta-analyses as criteria for real sex differences in nonverbal behavior.

Consensual stereotype accuracy: Discrepancies. They reported results separately for male and female perceivers. In general, the meta-analyses showed that women were nonverbally warmer (e.g., smiling more, interacting more closely) and more skilled (e.g., better face recognition, more decoding skill) than were men—and both men and women recognized these differences. Women were accurate for nine of the behaviors, they exaggerated sex differences on five behaviors, they underestimated sex differences on two behaviors, and there was one reversal (seeing a sex difference opposite to the real difference).

Men's stereotypes were accurate on 11 behaviors and they had 6 near misses (3 of which underestimated real differences, 1 of which exaggerated real differences, and 2 of which were reversals). They had no inaccuracies. Across the set of results, men tended to believe that sex differences were smaller than women believed them to be (reflected in men's slightly greater tendency to underestimate than exaggerate real differences and in women's slightly greater tendency to exaggerate real differences).

Both sexes also expressed gender-centric beliefs (beliefs favoring their own gender). When women exaggerated real differences, they did so by boosting their perception of women more than by derogating men. When men underestimated real differences, they did so more by boosting their perception of men than by derogating their perception of women. In other words, even though both men and women were quite accurate, when they made errors, those of women tended to be flattering toward women, and those of men tended to be flattering toward men.

Consensual stereotype accuracy: Correspondence with real differences. Although Briton and Hall (1995) did not report individual-level correlations, they did report correlations that reflect the accuracy of consensual stereotypes. How well did the average perceived sex difference correspond with the real differences obtained in the meta-analyses? Quite well indeed. Those correlations were .68 for men and .74 for women.

Conclusion. The evidence regarding consensual stereotypes demonstrated impressively high levels of accuracy among both men and women, both of whom hit bull's eyes on a majority of the sex differences. Another interesting aspect of their results was the demonstration of small in-group bias right alongside this evidence of accuracy. This is consistent with one of the major points of Chapter 10—that experimental demonstrations of bias do not preclude accuracy in real life because bias and accuracy can and often do exist side-by-side. This pattern is also broadly consistent with one of the main themes of this book—although biases are real, social beliefs are often more accurate than biased.

Of course, the study has some important limitations. Most important, it provided no information about individual-level stereotypes, which, most likely, were not as accurate as the consensual stereotypes (the reasons for this will be discussed in the next chapter). Furthermore, like many studies of stereotypes, its sample was limited to college students.

HALL AND CARTER (1999): INDIVIDUAL DIFFERENCES IN THE ACCURACY OF GENDER STEREOTYPES

Hall and Carter (1999) also compared college students' (over 700 of them) beliefs about the size of sex differences to results obtained in meta-analyses, with several new twists. One new twist was that, instead of merely studying 15 or 17 characteristics, Hall and Carter (1999) studied beliefs about 77 different traits and behaviors, organized into the following five categories: nonverbal communication (e.g., smiles, restlessness), cognitive performance (e.g., performance on standardized verbal and math tests), cognitive attitudes (e.g., confidence in math and science), personality (e.g., impulsiveness, trusting), and small group or organizational behavior (e.g., leadership effectiveness, persuadable). They examined both consensual and individual stereotypes and also examined whether a variety of individual differences predicted differences in stereotype accuracy. As such, it constitutes the most comprehensive study of the accuracy of sex stereotypes in the scientific literature.

Although they reported their results separately for men and women, in this study, the patterns of accuracy were nearly identical. Therefore, my summary of their results ignores the gender of the perceiver. They did not report discrepancy scores.

Consensual stereotype accuracy: Correspondence with real differences. Like Swim (1994) and Briton and Hall (1995), Hall and Carter (1999) assessed the accuracy of consensual stereotypes by correlating the average perceived sex difference with the real difference (as indicated by the meta-analyses). They found very high accuracy of consensual stereotypes—beliefs about men and women correlated .66 to 0.94 with real differences (and averaged .79, see their Table 3).

Personal stereotype accuracy: Correspondence with real differences. Next they assessed the extent to which their perceivers' personal stereotypes corresponded with the real characteristics of men and women. Separate correlations were computed for each perceiver, and then the median and mean correlations for the whole sample were summarized (in their Table 5). Some people were almost completely inaccurate (their perceived/actual correlations were near zero); others were highly accurate (their perceived/actual correlations were almost .7). The average of these individual-level correlations was .43. Most people were fairly accurate.

Who was more (in)accurate? The classic view holds that stereotypes are inherently evil, inaccurate, unjustified, rigid, and/or irrational (see Chapter 15). Although, by now, it should be clear that this classic view is, or at least should be, near death, perhaps it is possible to

revive it in a more limited or circumscribed form. Perhaps stereotypes are most likely to be inaccurate and not based in reality when the perceiver is prejudiced, predisposed toward exploiting people, or prone to extreme and rigid thinking. Now, not everyone fits this nasty profile, but, if you do, perhaps your stereotypes are not quite as accurate as everyone else's. Hall and Carter (1999) provided analyses capable of testing this idea.

First, they found only the barest hint of support for the idea that more prejudiced people hold more inaccurate stereotypes. They administered Glick and Fiske's (1996) Ambivalent Sexism Inventory, which has two subscales. The benevolent sexism subscale assesses the tendency to view women in overly positive, idealized ways; the hostile sexism subscale assesses the tendency to view women in overly negative, derogatory ways. Hostile sexism scores, however, did not predict stereotype accuracy for either men or women (this pattern is hard to find in the original article, because it only appears in their footnote 1 on p. 352). Thus, the stereotypes held by people who were prejudiced were just as accurate as stereotypes held by everyone else.

Benevolent sexism scores were unrelated to men's accuracy. Women scoring higher on benevolent sexism, however, were somewhat less accurate. Because benevolent sexism taps an overly idealized view of women (Glick & Fiske's [1996] interpretations of this subscale as assessing a form of prejudice notwithstanding), this may reflect an in-group bias effect. Women viewing women in an overly idealized manner (agreeing with items emphasizing the purity of women and the appropriateness of putting them on a pedestal) seems to me to be awfully close to viewing women more favorably than they deserve. If so, this may help explain why women scoring high on "benevolent sexism" held less accurate beliefs about men and women and their differences.

They also administered the Right-Wing Authoritarianism scale (Altemeyer, 1981), which is supposed to measure attitudes predisposing one toward support for fascist policies. As such, one might expect it to predict highly inaccurate stereotypes. It didn't (this, too, can only be found in footnote 1). So far, then, the study has not provided much support for the idea that prejudiced people hold less accurate stereotypes.

This idea, however, received more support from their analyses of social dominance (a scale measuring support for exploiting and oppressing others). Those hell-bent on taking advantage of other people did indeed hold less accurate stereotypes. Interestingly, however, social dominance predicted lower accuracy more strongly among women than among men.

They also administered a Universalism scale, which assessed the extent to which perceivers believed all people are essentially the same. The more people believed in universalism, the more accurate were their stereotypes. This result seems to be inconsistent with Wolsko et al. (2000), who found that people who adopted a "color blind" (emphasizing how we are all the same) perspective were less accurate than those who adopted a "multicultural" (emphasizing the importance of differences) perspective. I currently have no explanation for this inconsistency between studies, and it is interesting enough to warrant further research.

Hall and Carter's (1999) results regarding social sensitivity were also broadly consistent with the more modern view of stereotypes as reflecting real group differences. Specifically, they also assessed people's beliefs about how in touch they are with their social environment and their ability to judge others' nonverbal cues. People high on these measures (more in touch, good judges of nonverbals) held more accurate beliefs about sex differences.

These results, then, are important because they demonstrate that people who are generally more sensitive to others (widely viewed as mostly a good thing by psychologists) are also

more capable of making valid judgments about bona fide ways in which men and women differ. Of course, this makes sense because (1) if men and women often, on average, differ, as the meta-analyses show they do, then (2) people most in touch with what other people are like will also be most likely to detect and perceive actual sex differences. This pattern is broadly consistent with one of the main themes of this chapter—beliefs about group differences (stereotypes) reflect social reality more than they reflect bias, subjectivity, and distortion.

DO PEOPLE KNOW HOW MEN AND WOMEN ARE DISTRIBUTED INTO DIFFERENT OCCUPATIONS?

Five studies reported in two separate publications (McCauley & Thangavelu, 1991; McCauley, Thangavelu, & Rozin, 1988) assessed the accuracy of people's beliefs about the distributions of men and women into different occupations. The five studies examined a variety of occupations (doctor, nurse, lawyer, engineer, secretary, etc.) and compared people's beliefs to the actual distributions as indicated in the U.S. Census. There were a total of 521 participants across the five studies, which included students attending several different colleges, high school students, and railway commuters.

Consensual stereotype accuracy: Discrepancies. Discrepancy analyses showed that people were highly, though not perfectly, accurate. Across the five studies there were a total of 90 sex distribution judgments. Consensual stereotypes hit the bull's eye 56 times. Of the remaining 34 judgments, 28 were near misses (26 of these underestimated real differences, 2 exaggerated real differences), and 6 were inaccurate (all 6 underestimating real differences).

Consensual stereotype accuracy: Correspondence with real differences. How well did the beliefs about distributions correspond to the actual ones? The correlation of the average of the estimates of the proportion of women with the U.S. Census data was .94 to .98 (proportion of men is, of course, 100 minus the proportion of women, so there is no need to compute a separate correlation for proportion of men, because it would be identical). This level of accuracy is among the largest effects ever obtained, not just in stereotype research, not just in social psychology, but in all of the social sciences.

Conclusion. One potential criticism of this study is that estimating the distribution of men and women in different occupations is a relatively easy task. Maybe it is. But those defining stereotypes as inaccurate or who emphasize inaccuracy have never qualified their claims along the following lines: "Stereotypes are generally inaccurate, but they are often accurate for many judgments that are easy to make." Instead, stereotypes are just blanketly condemned for their inaccuracy, leaving no exceptions, even for easy tasks. Indeed, stereotypes, especially consensual stereotypes, are far more likely to be condemned as irrational mass-cultural myth than to be held up as an example of extraordinary social acuity under any conditions, even easy ones (e.g., Jost & Banaji, 1994; Katz & Braly, 1933).

Regardless, this is a fairly easy task. If people are reasonably in touch with reality, then they should be highly accurate on an easy task. They were.

A major strength of these studies is that they included noncollege as well as college samples and, indeed, showed the same pattern of high consensual accuracy accompanied by a far greater tendency to underestimate than overestimate real differences among all perceiver groups. A major limitation is that they only reported results for consensual stereotypes; no analyses addressed the accuracy of personal stereotypes.

ACCURACY IN PERCEIVING THE SEX DISTRIBUTION INTO OCCUPATIONS REPLICATED

This general pattern has been replicated in a more recent study of 189 introductory psychology students' beliefs about the distribution of men and women into 80 different occupational categories (Cejka & Eagly, 1999). These perceivers' beliefs were then compared to the actual sex distribution into these occupations according to the U.S. Census.

Consensual stereotype accuracy: Discrepancies. Because they did not report results separately for the 80 occupations, I cannot provide a specific breakdown of how often people were accurate or over- or underestimated real sex differences in distributions. However, they did report overall results separately for female-dominated and male-dominated occupations (averaging over all respondents and occupations). There was a general tendency to underestimate the real sex difference in the distribution into different occupations (underestimates of 9.3% and 17.1%, respectively, for male- and female-dominated occupations), which of course means that there was no general tendency to exaggerate real differences.

These figures suggest that many of the judgments would be considered accurate or near misses. It is, however, impossible to reach any clear conclusion about the accuracy of people's perceptions of distributions into specific occupations, because wild inaccuracies in opposite directions could cancel one another out (e.g., underestimating the difference by 40% for one occupation and overestimating by 30% for another means that, *on average*, people underestimated the real sex difference by 5%).

Consensual stereotype accuracy: Correspondence with real differences. The correlation of the average estimate of the proportion of women in each occupation with the actual proportion of women in each occupation was .91 (the proportion of men is a mirror image [% men = 100 - % women], so the correlation would be identical).

Conclusion. Assessing the accuracy of people's stereotypes was only a minor feature of this study, and, as a result, the evidence it provided is fairly sketchy. Nonetheless, it replicated the work by McCauley and colleagues demonstrating that people's stereotypes correspond very closely to actual occupational distributions and are far more likely to underestimate than exaggerate the real sex distribution into occupations. Cejka and Eagly (1999) suggest that this might reflect a "contraction bias"—a tendency to avoid extreme judgments. So, if people rarely estimate more than a 90% to 10% difference in sex distribution, and several jobs are 95% to 5% or more extreme, they will appear to underestimate real sex differences. Another possibility, however, is that people overestimate how egalitarian our society has become, so that they tend to guess that the sex distribution is more equal than it really is (these explanations, furthermore, are not mutually exclusive—both could be true to some degree). Currently, however, there is no scientific basis for determining whether either of these possibilities explains the pervasive tendency for people to *underestimate* the sex differences in distributions into occupations.

SEX STEREOTYPING OF MAJOR DISTRIBUTION AND ACADEMIC ACHIEVEMENT

Beyer (1999) examined the accuracy of 265 college students' beliefs about the distribution of men and women into 12 different majors (English, psychology, art, biology, etc.) at their college.

Over 75% of the students with a declared major were enrolled in these majors, so this was a fairly comprehensive list. These estimates were then compared to the actual distribution of males and females into the different majors as indicated by college records. She also assessed the accuracy of these students' beliefs about the GPAs of men and women in these various majors by comparing those beliefs to college records.

Consensual stereotype accuracy: Discrepancies. Because discrepancy scores were reported separately for male and female perceivers, there were a total of 24 judgments (12 majors by 2 sexes). Of these, 13 were accurate, 10 were near misses, and 1 was inaccurate. Of the 11 inaccuracies (including near misses), 6 underestimated real differences, 1 exaggerated real differences, and there were 4 reversals (both men and women erroneously believed that a majority of biology and business majors were men when, in fact, a majority in both cases were women). Overall, therefore, these results are broadly consistent with those of the research on sex stereotypes regarding occupational distributions, in showing substantial accuracy and a greater tendency to underestimate than exaggerate the real sex difference in major distribution.

In general, people also generally underestimated the proportion of women majors (some degree of underestimation occurred on 19 of 24 judgments); across all majors, people underestimated the proportion of women by about 7%. This is a deliciously ambiguous result. It could reflect prejudice against women—the belief that women are not as smart or as ambitious as they really are. If people underestimate women's intelligence and ambitiousness, they may also underestimate their likelihood of attending college.

However, it could also reflect the “common knowledge” that women are oppressed. If there is widespread belief that society advantages men over women in school, then “of course” there should be more men than women in college. Given that there are actually more women than men in college (both nationwide and in Beyer's data, about 55% of college students are women [Beyer, 1999]), a widespread belief that schools disadvantage girls could also lead people to assume that fewer women than men attend college—thereby producing widespread underestimation of the distribution of women across majors. I refer to this as the “belief that society oppresses women” explanation.

So, is it prejudice or a belief that society oppresses women? Beyer's data do not allow us to distinguish between these explanations in order to answer this question (although some of Beyer's additional data—soon to be discussed—as well as studies reviewed later in this chapter point more toward the “belief that society oppresses women” explanation than toward actual prejudice).

Beyer (1999) also compared her perceivers' estimates of men's and women's GPAs in each of the 12 majors (producing 48 comparisons: sex of perceiver by sex of target by the 12 majors). Overall, there was a general tendency to overestimate students' GPAs. This occurred in 47 of 48 judgments and, overall, averaged 0.26 of a GPA point (this is an “elevation” effect; see Chapter 12).

How accurate were the perceived *differences* between men and women's GPAs? Unfortunately, Beyer (1999) did not report standard deviations, so I cannot use my 0.25 SD criteria (and GPAs are not percentages, so I cannot use my 10% criteria). So, I will be conservative—if they get the real difference within 0.1 of a GPA point, I will call them accurate; if the perceived difference is more than 0.1 but no more than 0.2, I will call it a near miss; any more than that, I will call it inaccurate. This seems like a very high standard for accuracy, but let's see where it gets us.

There were 12 perceived differences, and Beyer (1999) reported the results separately for male and female perceivers. Male perceivers were accurate four times, had five near misses, and were substantially inaccurate three times. Female perceivers were accurate twice, had six near misses, and were substantially inaccurate four times.

The sources of the errors were quite interesting. Both sexes tended to overestimate males' GPAs more than females' GPAs. However, women did this more than did men. This could be a case where women are biased in favor of men more so than are men themselves. But I doubt it. First, it is an unusual pattern. Indeed, if this interpretation were true, this would be the only study of stereotype accuracy to demonstrate a pattern wherein an in-group's stereotypes are more biased against itself than against an out-group.

Another reason to suspect that the women's greater overestimation of men's GPA does not reflect a general bias in favor of men is that a general bias should manifest in all majors, not just masculine majors. This, however, did not happen. Instead, the most pronounced overestimation of men's GPAs occurred when women estimated men's GPAs in masculine majors (when, in fact, women had higher GPAs than men across the board, even in masculine majors).

Instead, the greater overestimation of men's GPAs across the board, but especially by women, may be more consistent with the "belief that society oppresses women" explanation. Women probably expect greater discrimination in historically male majors (engineering, math, the sciences) than in other neutral majors (such as history, sociology, art, nursing). If so, that belief may manifest as an overestimation of the effects of that favoritism—by overestimating men's GPAs. The idea that men do better than women in college, and especially (though not exclusively) in masculine majors, may be one of those pervasive cultural myths perpetuated by well-intentioned social activists and researchers (e.g., *How Schools Shortchange Girls*, American Association of University Women, 1992, a title I consider unintentionally ironic given that, on average, girls do better at every level of schooling than do boys).

Consensual stereotype accuracy: Correspondence with real differences. The correlation between mean perceived distribution and actual distribution into the different majors was .80 for male perceivers and .79 for female perceivers. For GPA, the consensual stereotype accuracy correlations were .35 when estimating the GPAs of males and .34 when estimating the GPAs of females.

Personal stereotype accuracy: Correspondence with real differences. Although Beyer (1999) did not assess the accuracy of personal stereotypes with discrepancy scores, she did report correlations between individual perceivers' beliefs and the criteria. When estimating the proportion of men and women in the various majors, the average correlations of individual estimates with the real distribution were .52 for women and .48 for men. Her perceivers, individually, were highly sensitive to the real sex differences in distributions into different majors.

Individual-level (personal stereotype) correlations also showed that beliefs about GPA were not very sensitive to actual GPAs. There was no sex of perceiver difference in these correlations, so Beyer (1999) reported them separately by sex of target. Those results showed that her perceivers were pretty clueless about differences in women's GPAs in the different majors (correlation of estimated and actual GPA for female targets was -0.04) and only slightly better when judging men's GPAs in the different majors (correlation of .22).

Conclusion. Beyer's (1999) study reconfirmed the predominant pattern from prior research demonstrating that people frequently underestimate real differences between groups. She showed that people were far more accurate in estimating the sex distribution into different majors than in estimating men's and women's GPAs. Even the GPA data, however, showed moderate accuracy in the consensual stereotypes, although little accuracy in the personal stereotypes. This is a very interesting finding for several reasons. First, it (like most of the other research reviewed in this chapter) disconfirms perspectives emphasizing consensual/cultural stereotypes as false, shared cultural myths. The consensual stereotypes, as is typical, were more, not less, accurate than personal stereotypes. Second, one might be wondering, "How can the consensual stereotypes have any accuracy, when the individual stereotypes were highly inaccurate?" This issue will be considered in Chapter 19.

In addition, people underestimated the number of women in the various majors. They also incorrectly believed that women's GPAs were lower than men's, when, in fact, men's GPAs were lower. Although the possibility cannot be ruled out that the women in this study were more prejudiced against women than were the men, a more plausible explanation is that everyone, but especially women, believed that, at least as far as college is concerned, women are more oppressed than they really are.

The study also has two relatively minor limitations. Beyer (1999) did not report discrepancy score results for her perceivers' personal stereotypes, so we do not know how far from perfection most of them were. Also, she only studied a college sample, which is important because it provides no information about the accuracy of people not in college.

HOW WELL DO PEOPLE KNOW THE POLITICAL ATTITUDES OF MEN AND WOMEN?

Social psychologists frequently think of stereotypes in terms of personality traits and, occasionally, in terms of behavior or accomplishments. This has been reflected in nearly all of the stereotype accuracy studies reviewed thus far. Diekmann, Eagly, and Kulesa (2002), however, focused on a different domain—attitudes. Specifically, across three studies, they had over 600 college students estimate the extent to which men and women supported various attitude and policy positions. For some of the positions, support by women was stereotypical (e.g., employers should offer paid time off to new parents); for others, support by men was stereotypical (e.g., favoring less government regulation of business); and others were nonstereotypic of either group (e.g., support for the United States taking an active part in world affairs). Two studies assessed the accuracy of sex stereotypes regarding 15 of these attitude questions; one study used 16 attitude questions. Perceptions of men's and women's attitudes were then compared to the actual attitudes, as indicated in the General Social Survey, which is a recurring nationally representative survey of Americans.

Consensual stereotype accuracy: Discrepancies. Unfortunately, Diekmann et al. (2002) did not report their results separately for the 15 or 16 attitude items; instead, they reported results averaged over all attitudes. They did, however, report results separately for men and women targets.

Therefore, there were a total of six consensual stereotype accuracy discrepancies reported—three studies each by men and women targets. All six were underestimates—that is, there was a general tendency to underestimate how much people supported the various positions.

However, this tendency was not very large; the consensual estimates underestimated support by only 2% to 8%.

Unfortunately, Diekman et al. (2002) did not directly assess whether people generally exaggerated real differences between groups. However, people most strongly underestimated the support of men for stereotypically female attitude positions. If we assume that, in general, women actually supported these positions more so than men (which cannot be determined from their data), then underestimating men's support would have the effect of exaggerating real differences between the groups.

Additional analyses provided further insight into how and why people went wrong. They administered another survey assessing the extent to which people believed each of the attitude positions produced positive and negative implications for women or men. The more positive the implications for women of a particular attitude position, the more people underestimated how much men agreed with that position. So, this may be a case where an underlying stereotype ("men are sexist"), which itself seems to reflect overestimates of oppression, seems to be undermining the accuracy of people's beliefs about men's specific attitudes.

Consensual stereotype accuracy: Correspondence with real differences. The correlation between people's consensual beliefs about men's and women's attitudes and their actual attitudes ranged from .53 to .80 and averaged .72 across the three studies. Once again, consensual stereotypes corresponded extraordinarily well with the groups' actual characteristics.

Personal stereotype accuracy: Discrepancies. Diekman et al. (2002) did not perform a full and clear assessment of the extent to which personal stereotypes deviated from perfection in the manner that, for example, Ashton and Esses (1999) did. They did, however, report the average "absolute" discrepancies for male and female targets for each of the three studies. These results refer to how far off from perfection people were, ignoring whether they over- or underestimated the real levels of support for the attitude or policy statement. For example, if Fred overestimated support by 10% and Beatrice underestimated support by 20%, their average absolute discrepancy would be 15%. This number, then, comes close to assessing the average accuracy of people's personal stereotypes, again, keeping in mind that they reported results averaged over all 15 or 16 attitude questions.

On average, people were off by 17% to 19% for both male and female targets (across all three studies). This might mean that they were wildly off on a few questions and pretty accurate on many or consistently off by about 15% to 20%—the data they reported do not allow us to identify the particular pattern of (in)accuracies that they found.

Personal stereotype accuracy: Correspondence with real differences. Diekman et al. (2002) also assessed, for each perceiver, the correlation of their beliefs about men's and women's support for the various attitude positions with men's and women's actual support for the various attitude positions. Their Table 1 reported the average of these personal stereotype correlations, separately for male and female targets by each of the three studies. Those correlations were moderately high, ranging from .34 to .58 and averaging .50. People were, apparently, quite (though not perfectly) sensitive to differences in the extent to which men and women supported the various attitude positions.

Conclusion. Overall, consensual stereotypes (discrepancies and correspondence) about men's and women's attitudes were extraordinarily accurate. Personal stereotypes also corresponded well with real differences, even though, in absolute terms, they were somewhat discrepant from the criteria.

These results are amazing or, at least, should be amazing to anyone familiar with the traditional social psychological discourse emphasizing the power of error and bias to run rampant in the types of “ambiguous” social situations that are supposedly so common (e.g., Darley & Fazio, 1980; Gilbert, 1995; Jones, 1986, 1990; Nisbett & Ross, 1980; Ross & Nisbett, 1991). Distribution into different jobs, nonverbal behavior, and the like are all observable or objective; attitudes are the type of vague and fuzzy characteristic that should, according to conventional social psychological wisdom, limit accuracy and allow people’s tendency toward bias and distortion to have maximum effect. And yet, even when judging such a fuzzy and ambiguous characteristic as an attitude, people’s stereotypes were often moderately, and sometimes highly, accurate.

These results nicely complement those of Beyer (1999), suggesting that people overestimate “oppression.” The results of Diekmann et al. (2002) show that people assume men are more sexist (less supportive of positions supporting women) than they really are. Such a general assumption, then, would explain why people’s stereotypes of men’s support for women’s issues consistently and most strongly underestimate men’s actual support.

Of course, this study, too, had important limitations. It did not provide very detailed information about accuracy, instead reporting results averaged over all 15 or 16 attitude positions. This rendered it impossible to determine how frequently people’s stereotypes corresponded to men’s and women’s stated attitudes. And, like most other studies, its perceivers were exclusively college students, so we really do not know whether adults out in the world of work, raising families, etc., hold more or less accurate stereotypes about men’s and women’s attitudes.

Other Stereotypes

When people think about stereotypes, race, ethnicity, and gender are probably among the types that most readily come to mind. Nonetheless, stereotypes are not restricted to groups for whom politicized issues of oppression, discrimination, injustice, and inequality are salient. Stereotypes are beliefs about groups, and there are many types of groups other than race, ethnicity, and gender. Furthermore, social science perspectives on “stereotypes” rarely limit their assumptions of inaccuracy or exaggeration to demographic stereotypes. I have never read any perspective on stereotypes claiming anything like “race and gender stereotypes are widely inaccurate, but occupational, regional, and political stereotypes are often nicely in touch with reality.” And, occasionally, the accuracy of these other stereotypes has been assessed.

Table 17–3 summarizes the results of all studies that have assessed the accuracy of stereotypes regarding groups other than race, ethnicity, or sex that I could find that met the criteria for inclusion described in Chapter 16. Each study is described next.

OCCUPATIONAL AND COLLEGE MAJOR STEREOTYPES

Beliefs about pay. The Cejka and Eagly (1999) study described previously in the section on the accuracy of sex stereotypes also included one result relevant to occupational stereotypes. Their participants were asked to estimate the average wage in the various occupations they

TABLE 17-3

The Accuracy of Other Stereotypes

Study and Stereotype	Perceivers	Criterion	Predominant Pattern of Discrepancies ^a	Individual Correlations (Personal Stereotype Accuracy)	Aggregate Correlations (Consensual Stereotype Accuracy)
Judd et al. (1991): beliefs about engineering and business majors at University of Colorado	116 University of Colorado business and engineering majors (58 each) randomly selected	Self-reports of those randomly selected 116 business and engineering majors	Accuracy and exaggeration	.63 ^b	Not available
Judd and Park (1993): Democrats' and Republicans' beliefs about one another's political attitudes	An unspecified number of people randomly surveyed as part of the 1976 National Election Study	Self-reported attitudes of self-identified Democrats and Republicans	Accuracy ^c and "liberalism bias": the tendency to overestimate the liberalism of members of both parties	.25 ^b	Not available
Cejka and Eagly (1999): wages in different occupations	189 introductory psychology students	Census data on 80 occupations	"Contraction bias": the tendency to overestimate wages in low-wage jobs and underestimate wages in high- wage jobs	Not available	.94

(Continued)

TABLE 17-3

The Accuracy of Other Stereotypes (Continued)					
Study and Stereotype	Perceivers	Criterion	Predominant Pattern of Discrepancies ^a	Individual Correlations (Personal Stereotype Accuracy)	Aggregate Correlations (Consensual Stereotype Accuracy)
Ryan and Bogart (2001): beliefs sorority members hold about their own and other sororities	136–181 sorority members (attrition over time due to graduation and dropping out of school or sorority)	Self-reports of 85%–100% of the full members (not including pledges and new initiates) of each sorority	Accuracy when perceiving their own sorority; exaggeration of stereotypicality when perceiving other sororities	When perceiving their own sorority: .52 ^b When perceiving other sororities: .39 ^b	Not available
Clabaugh and Morling (2004): beliefs about the psychological characteristics of ballet dancers and modern dancers	175 ballet dancers, modern dancers, and psychology students	Self-reports of the ballet dancers and modern dancers	Accuracy	<i>Perceiving differences between groups:</i> Ballet perceivers: .67 Modern dance perceivers: .71 Psych student perceivers: .62	<i>Perceiving differences between groups^d:</i> Ballet perceivers: .83 Modern dance perceivers: .90 Psych student perceivers: .79

Perceiving differences
across traits within
target groups
Ballet perceivers: .59^b
Modern dance
perceivers: .67^b
Psych student
perceivers: .45^b

^a Except where otherwise stated, all discrepancy results occur at the consensual level. Except where otherwise noted, (1) accuracy means within 10% of the real percentage or within 0.25 of a standard deviation, (2) exaggeration means that the perceived differences between groups exceeded the group differences on the criteria, and (3) underestimation means that the perceived differences between 52 groups was smaller than the group differences on the criteria. Only one word is entered in this column when one pattern (e.g., “accuracy”) occurred for a majority of results reported. When there was no majority (or the majority could not be determined from their data), the top two results, in order of frequency (most frequent first), are reported here.

^b For simplicity, if the study reported more than one individual-level (average) correlation, I have simply averaged all their correlations together to give an overall sense of the degree of accuracy.

^c Neither percentages nor standard deviations were reported. I characterize the main results of their discrepancy analyses as “accurate” because seven of eight mean discrepancies are all less than 1 scale point (on a 7-point scale).

^d These correlations do not appear in the original article, but are computable from data that was reported.

examined, thereby assessing one aspect of occupational stereotypes. They reported two primary results: (1) high accuracy of the consensual/aggregate stereotype (the correlation between perceived and actual mean wages was .94) and (2) a “contraction bias” whereby people tended to underestimate wages in high-paying jobs and overestimate wages in low-paying jobs (although the extent of this bias is unclear because they did not report the data). They did not provide any data about personal stereotypes regarding wages in different occupations.

Beliefs about dancers. One of the most unique studies of stereotype accuracy was conducted by Clabaugh and Morling (2004), which investigated the accuracy of stereotypes regarding modern dancers and ballet dancers. Because of the paucity of research on occupational stereotypes and because of the quirkiness of the topic, I have decided to make an exception to my decision not to report studies that used criteria that were mismatched to the stereotype. This study asked perceivers to rate modern dancers and ballet dancers in general, but then used the self-reports of the haphazard samples of dancers in their study as criteria.

Of course, both the use of a haphazard sample as criteria and self-reports are limitations of the study. It is important to keep in mind, however, that both limitations would likely *decrease* evidence of accuracy (because of the mismatch between the stereotype assessed and the criterion sample and because of the potential inaccuracies in self-reports). Therefore, the true accuracy of perceivers in this study probably exceeds whatever empirical evidence of accuracy that they obtained.

This study examined the stereotypes held by modern and ballet dancers attending a professional dance camp and by a sample of introductory psychology students (total N for all three groups was 175) regarding modern and ballet dancers. Specifically, they assessed people’s beliefs about the self-esteem, body image, physical condition, fear of negative evaluation, need for structure, and need for control regarding the different dancers and used the dancers’ self-reports on these same items as criteria. Unfortunately, the questionnaire assessed beliefs about modern and ballet dancers in general, rather than about those attending the camp, thereby creating the mismatch between the stereotype assessed and the criteria.

Clabaugh and Morling (2004) did not report the mean perceptions of *the level of each trait* for each group. However, they did report the mean *differences* between groups in self-reports and the mean perceived differences between the groups. Therefore, the extent to which the consensual stereotype regarding *differences* corresponded to the real differences could be computed from their data (they did not report this correlation). These consensual beliefs about group differences corresponded extremely well with differences in the dancers’ self-reports. Consensual stereotypes correlated with the self-reports .83, .90, and .79 for perceivers who were, respectively, ballet dancers, modern dancers, or introductory psychology students.

They also assessed the extent to which personal stereotypes corresponded with the dancers’ self-reports. One set of analyses assessed how sensitive people were to different levels of each characteristic *within* each group (e.g., how sensitive are people to differences in the body image, self-esteem, etc., of ballet dancers?). For each perceiver, Clabaugh and Morling (2004) computed the correlation of their perceived level of each characteristic with the self-reported mean level. Among individual psychology students, these correlations indicated only modest sensitivity to variations in the traits of ballet dancers (average correlation between beliefs and criteria was .23). Among all other combinations of perceiver group

(ballet, modern, intro psych student) and target group (ballet and modern), these correlations were substantial (the average correlations ranging from .48 to .63).

They also assessed people's sensitivity to *differences* between ballet and modern dancers (how well do perceived differences correspond to the self-reported differences?). Again, these were personal stereotypes, because they computed these correlations for each perceiver. These average correlations were strikingly high: .67 for ballet perceivers, .71 for modern dance perceivers, and .62 for the introductory psychology students.

Discrepancy scores showed the now familiar pattern of high accuracy and some systematic error. At the consensual stereotype (aggregate) level, 11 of 18 perceived differences (6 traits by 3 groups of perceivers) were accurate, 6 were near misses, and 1 was inaccurate. Of the seven inaccuracies (including the six near misses), four exaggerated the real difference, one underestimated the real difference, and there were two reversals (both groups of dancers believed that modern dancers had higher "body esteem," although ballet dancers reported higher body esteem than did modern dancers). They did not report personal discrepancy scores.

Overall, therefore, this study provided strong evidence of accuracy nearly across the board (discrepancies, correspondences, personal and consensual stereotypes). One could consider the quirkiness of the study to be a limitation. After all, it has not addressed any of the major stereotypes traditionally believed to play a role in oppression and inequality. Indeed, stereotyping of dancers is not exactly a hot social issue. I, however, consider this quirkiness to be a scientific strength, even if it is not a political strength. When highly similar patterns (in this case, of moderately high individual accuracy, very high consensual accuracy, and some systematic discrepancy) occurs with both common and quirky stereotypes, we become more able to draw conclusions about the accuracy and inaccuracy of stereotypes in general, and not just the handful of hot-button stereotypes that have received most of the attention.

Beliefs about engineering and business majors. Judd, Ryan, and Park (1991) examined the beliefs about engineering and business students held by random samples of 58 engineering and 58 business students at the University of Colorado. These students rated each group on eight trait and attitudinal items, four of which were stereotypical of business students and counterstereotypical of engineering students (e.g., extroverted) and four of which were stereotypical of engineering students and counterstereotypical of business students (e.g., "One of my favorite pastimes is solving brain teasers such as Rubik's Cube"). These same students' self-reports on these same questions were the criteria for assessing the accuracy of the stereotypes.

The discrepancy analyses in Judd et al. (1991) were reported for the whole set of perceivers and, therefore, constituted consensual stereotype assessment. They did not report data assessing the accuracy of personal stereotypes as discrepancies. Unfortunately, Judd, et al. (1991) did not report results separately for each of the eight items, so I cannot tell you how often the students were accurate. Instead, their Table 7 reported the overall average tendency to over- or underestimate stereotypicality across the eight items (underestimating counterstereotypicality was treated as if it was essentially the same thing as overestimating stereotypicality). They reported four overall mean tendencies to over- or underestimate stereotypicality (students from two majors judging students from two majors averaged over all eight items). Although all were positive (indicating some tendency to overestimate stereotypicality—i.e., exaggerate real differences), three of the four were under 10% and the fourth was 18%, strongly suggesting that there was considerable accuracy here. They did not report results that would permit assessment of the extent to which the consensual stereotypes corresponded (correlated) with their criteria.

They did, however, assess the extent to which personal stereotypes corresponded to their criteria. Specifically, they examined how well people's beliefs about how the groups varied from trait to trait corresponded with how each group actually varied from trait to trait. This, therefore, assesses correspondence of belief to criteria *within* each group; it does not assess how well people's beliefs about differences between the groups corresponded to the actual differences between the groups. These correlations ranged from .43 to .78 and averaged .63, indicating that people were highly sensitive to differences on the eight criteria among business majors and among engineering majors. They also found that people's personal stereotypes of their own group corresponded with criteria somewhat better (.71) than did people's stereotypes of the other group (.54), although both correlations indicate substantial accuracy.

POLITICAL STEREOTYPES

Judd and Park (1993) examined the accuracy of Democrats' and Republicans' perceptions of one another based on the 1976 National Election Study, which surveyed a representative sample of Americans. They compared people's beliefs to 10 self-reported attitudinal positions of people identifying themselves as Democrats and Republicans (e.g., support for school busing to achieve racial integration, taxes, etc.). These positions were assessed using 1-to-7-point scales, so discrepancy score analyses all involved comparison of belief to self-report on these scales.

They did not report these discrepancy scores for each individual, so their results focused on the consensual stereotype. The discrepancy score results for consensual stereotypes were quite interesting: (1) In general, both Democrats and Republicans overestimated both parties' liberalness; (2) Republicans did this generally more so than did Democrats; and (3) the liberalness of Democratic targets was overestimated more than the liberalness of Republican targets. This is tangential, but I find this interesting because it may help explain the conservative shift in the country that took place from about 1968 to 2000. If both parties are seen as more liberal than they really are, moderates will shift their support to Republicans, because Democrats (who are, on average, left of the political center) will be seen as even farther to the left and because Republicans (who actually are, on average, right of the political center) will be seen as more moderate than they really are. Thus, Republican candidates would likely receive support from both conservatives and moderates, and many Democratic candidates would be left primarily with liberals.

Returning to stereotypes, these results are also consistent with a sort of tortured version of the exaggeration hypothesis. If Democrats are seen as much more liberal than they really are, and if Republicans are seen as only slightly more liberal than they really are, the perceived political differences between Democrats and Republicans will be seen as greater than they really are (despite the displacement of both parties to the left).

Unfortunately, the discrepancy scores were not reported on a percentage scale and Judd and Park (1993) did not report standard deviations, so I cannot evaluate the accuracy of their perceivers' judgments using either my "within 10%" or "within .25 SD" standards. However, their Table 5, which reports eight separate discrepancy scores (strong and weak Democrats and Republicans judging Democrats and Republicans) shows that seven of the eight discrepancy scores (inaccuracy) were modest—less than a single point (on the 7-point scale). Only in the case of strong Republicans judging Democrats did the discrepancy score exceed 1. It seems, then, that there was likely considerable accuracy here by any reasonable absolute standard. Unfortunately, they did not report personal discrepancy scores.

They did, however, assess how sensitive individual perceivers were to variability in support for the various positions among Democrats and among Republicans (so, these analyses are only performed within each perceiver group; they did not assess the accuracy of people's perceptions of differences *between* each group). These average correlations ranged from .11 to .44 and the overall average was about .25. Furthermore, they found that these correlations were higher when people judged the attitudes of members of their own party (correlations of about .2 to .4) than when people judged the attitudes of the other party (correlations of about .1 to .2). Like Judd et al. (1991), they found that people's personal stereotypes corresponded more with the characteristics of their own group than with an out-group.

These correlations indicate some of the lowest levels of accuracy in any of the studies I have yet found on the accuracy of stereotypes. Experts and pundits have long bemoaned the political ignorance and apathy of the general population. These results indicate that people do indeed seem to know a lot less about politics than they know about other spheres of life, including gender-, ethnicity-, and occupational-related characteristics and behaviors.

On the other hand, we (at least, those of us who are psychologists) probably do not want to too loudly condemn people's "inaccuracy" when their beliefs "only" correlate .25 with criteria. Correlations of .25 fall into Cohen's (1988) moderate range and can also be viewed as people being right about 63% of the time, per Rosenthal's binomial effect size display. If you are a psychologist tempted to view .25 as worthy of derision, please keep in mind that you are also implicitly derogating your entire discipline as "inaccurate," because .25 is larger than a majority of effects obtained in psychological research (Hemphill, 2003; Richard, Bond, & Stokes-Zoota, 2003). What's good for the goose is good for the gander; what's valid for the scientist is also valid for the layperson. Or, at least, it should be.

SORORITIES

Ryan and Bogart (2001) examined the accuracy of the beliefs about their own and other sororities held by 84 new members of four sororities. Extensive pilot testing and a prior study had identified attributes that were stereotypic of one sorority and counterstereotypic of one sorority. Attributes included traits (e.g., competitive), behaviors (e.g., challenges authority), and attitudes (e.g., believes Jewish holidays should be observed by the university). They also surveyed all current members of the sororities (not including the new members) to obtain self-reports regarding these attributes and used these aggregated self-reports as criteria.

Following Judd and Park's (1993) approach (described in Chapter 12), Ryan and Bogart (2001) examined discrepancy scores after averaging perceived-actual differences (for each perceiver) across all judgments of a group. As a result, they did not report perceptions and actual scores on criteria (or their differences) for individual items; they simply reported overall (averaged) discrepancies. Unfortunately, this means that we are unable to determine how frequently people's stereotypes were accurate and inaccurate; all that we can know is the average levels of discrepancies.

These overall average discrepancies showed that people generally underestimated both stereotypic and counterstereotypic attributes, but that they underestimated counterstereotypic attributes more. This means that people saw the groups as more stereotypic than indicated by the criteria. Furthermore, this pattern of overestimating stereotypicality was more extreme when judging an out-group (another sorority) than when judging the in-group

(one's own sorority). This also means, therefore, that people exaggerated the real differences between the sororities.

Nonetheless, at the consensual level (keeping in mind that their reported numbers also collapse over all judgments), there was also considerable evidence of accuracy. Reported results included consensual discrepancies separated by in-group/out-group, stereotypic/counterstereotypic, and positive/negative at four different time points. So they reported a total of 32 consensual discrepancies. Of these, 13 were accurate, 14 were near misses, and 5 were inaccurate. Of the 19 inaccuracies (including near misses), 16 underestimated the level of the traits in the target group, and 3 overestimated the level of the traits in the target group.

They also reported the difference between stereotypic and counterstereotypic discrepancies separately for in-groups and out-groups at each of the four time points (eight differences total). When judging their in-group, all four of these stereotypic minus counterstereotypic discrepancies were accurate. For the out-group, none was accurate: two were inaccurate, two were near misses, and all four overestimated stereotypicality. This result is consistent with the exaggeration hypothesis (overestimating stereotypicality in the Judd & Park, 1993, componential system constitutes exaggeration—see Chapter 12).

Like Dickman et al. (2002), Ryan and Bogart (2001) also reported absolute discrepancies (treating both underestimating by 20% and overestimating by 20% as the same). These are roughly interpretable as personal stereotype discrepancy scores, because they represent the average total amount each individual was wrong (regardless of direction of the wrongness). Unfortunately, however, they did not report these results separately for each attribute. Instead, they reported the total discrepancies for several attributes at a time (separately for in-group/out-group, stereotypic/counterstereotypic, and positive/negative attributes, at each of four separate times—32 absolute discrepancies, total). Therefore, we only know the average discrepancy for sets of attributes, rather than for each attribute. These ranged from about 14% to 36%, indicating substantial inaccuracy (and, again, people were more inaccurate judging other sororities than judging their own).

They also did not report correlations representing the correspondence of the consensual stereotypes with criteria (or data from which those correlations could be computed). They did, however, report correlations representing the correspondence of the personal stereotypes with criteria. Specifically, they reported correlations between people's beliefs about the levels of each trait *within* a sorority and the average self-reported trait levels for that sorority. Overall, those correlations averaged .46, although people were more accurate judging the traits of their own sorority (.52) than of other sororities (.39). They did not report correlations reflecting the accuracy of people's beliefs about differences between the different sororities. Personal stereotypes, then, were reasonably sensitive to how the different sororities (on average) varied across the criteria.

DO THEY ALL LOOK ALIKE TO YOU? BELIEFS ABOUT DISPERSION

One of the classic accusations leveled against stereotypes is that they are “overgeneralized”—people see groups and their members as more similar to one another than is justified. Often when this is discussed, little distinction is made between a *generalization* and an *overgeneralization*. As discussed throughout this book, a generalization can still be valid without being perfect. “Birds fly” is a reasonable and valid generalization, despite the existence of penguins

and ostriches. “All birds fly,” however, would be an overgeneralization—a belief that treats birds as more similar to one another than they really are.

Overgeneralization, however, can be separated into two issues: (1) Is there a general tendency to perceive groups as less variable than they really are (I refer to this as “overgeneralization”)?; and (2) Are people more likely to underestimate variability when thinking about out-groups than when thinking about their in-groups (this pattern is often referred to as “out-group homogeneity” in the literature)?

Much of the research addressing these questions has come from the Judd/Park/Ryan group, who often, but not always, find evidence of both overgeneralization and out-group homogeneity (Judd & Park, 1993; Judd et al., 1991; Ryan, 1996). Other researchers, however, often have not found these patterns (e.g., Lee & Ottati, 1993; Linville, Fischer, & Salovey, 1989; Simon & Pettigrew, 1990). Although the reasons for this difference among research teams are not clear to me, the following conclusions appear warranted:

1. Overgeneralization and out-group homogeneity have been found enough to warrant continued study, especially of the conditions under which they are more or less likely to occur.
2. Neither has been found consistently enough to warrant inclusion as a defining or general feature of stereotypes.

WHAT ABOUT PERSON PERCEPTION?

OK, so the scientific evidence does not justify concluding that stereotypes are pervasively inaccurate; instead, stereotype accuracy is one of the largest effects in all of social psychology. Before discussing the implications of these results, it is necessary to address one more “yes, but.” Specifically, “Yes, but what is really important about stereotypes is how they lead to biased judgments regarding individuals.”

This type of “yes, but” argument is fairly common in modern perspectives on stereotypes (e.g., Fiske, 2004; Nelson, 2002; Schneider, 2004; Stangor, 1995) and, in essence, engages in a major tactical retreat from perspectives that have emphasized stereotype inaccuracy in the past. It grudgingly acknowledges that stereotypes are, in some sense, somewhat accurate for overall perceptions of groups but goes on to imply that this is not very important. Instead, according to this tactical retreat, what is important is how people perceive and behave toward individual members of different groups.

I doubt that most proponents of this view would characterize it as any sort of “retreat,” let alone a “tactical” one. Nonetheless, I do so characterize it. That is because this view at least acknowledges that stereotypes as perceptions of groups may indeed sometimes be pretty accurate. As such, it constitutes a serious retreat from the overwhelming emphasis on inaccuracy that has characterized most of the social science discourse on stereotypes.

At the same time, however, this “yes, but” can be viewed as “tactical” because it denigrates the importance of finding evidence of stereotype accuracy. As such, it allows the proponents of this view to maintain intact a view of stereotypes as “generalizations gone rotten” (Schneider, 2004)—not because they are inaccurate per se, but because of the allegedly awful and inappropriate ways people apply them when perceiving and judging individuals. *“Yes, stereotypes may not always be inaccurate, but let’s get back to bias. . . .”*

I find this type of perspective deliciously ironic. After 90 years of psychologists proclaiming to the world the inaccuracy of stereotypes in books, scientific articles, chapters, and reviews, when the hard evidence finally comes in indicating that stereotypes are often pretty accurate, some psychologists conclude that this is not very important. How can it be important to proclaim their inaccuracy but not their accuracy? The simplest answer is that this type of hypocritical stance reflects politics, not science. If the purpose of science is to serve political ends, such as combating inequality, then it is important to proclaim stereotype inaccuracy because it might help end inequality. Proclaiming stereotype accuracy, however, does not obviously help end inequality, so that, if the motivations are entirely political and not scientific, proclaiming it would understandably be viewed as entirely uninteresting and unimportant.

Regardless, I respectfully disagree. It seems to me terribly important to know whether people's beliefs about groups are largely in touch with reality or, instead, are largely irrational manifestations of bigotry, or even social constructions based on national or ethnic mythologies. Nonetheless, there is one aspect of this tactical treat with which I do agree—understanding whether and when stereotypes increase or decrease the accuracy of person perception is indeed an important issue. This chapter did not address that issue.

Whether or not stereotypes are “accurate” in the sense of being reasonably valid descriptions of broad group differences does not preclude the possibility that they frequently lead people astray when judging individuals. Even if Asian Americans, on average, really do have higher math scores and are more interested in engineering than other Americans, it is still unreasonable to assume that any given Asian American high school graduate is a math whiz hell-bent on a career in engineering, isn't it? Well, when phrased in that “let's present what people believe in as extreme and as exaggerated a way as possible so we can beat our breasts in righteous indignation about the invalid and presumptuous nature of their stereotypes” sort of way, even I have to agree that it is inappropriate to presume that any given Asian American student is a math whiz.

The issue is, however, do people actually think in that sort of extreme and exaggerated way? Or does this sort of breast beating simply set up a straw argument in order to look righteous when we knock it down (“it's just not true that all Asians are engineers!”). My claim that famous and influential social psychologists make this type of straw argument is itself most definitely *not* made of straw, as the following quotes quite clearly show:

“It is simply not true that all Germans are industrious or that all women are dependent and conforming” (Snyder, Tanke, & Berscheid, 1977, p. 657).

“Once an individual is classified as a member of a social group, perceptions of that group's average or reputed characteristics . . . are readily relied on by those doing the classifying. It then becomes more difficult for the classifier to respond to the other person's own particular characteristics, making accurate, differentiated, and unique impressions less likely. In such instances, people tend to perceive members of the other group as all alike or to expect them to be all alike, which they never are” (American Psychological Association, 1991, p. 1064).

Apparently, the consensus in the American Psychological Association (APA) was that, when people classify others into groups, they perceive them as all alike or expect them to be

all alike. Chapter 18 reviews the data and concludes that there is, in fact, little empirical support for such an extreme claim. The issues involved in understanding the role of stereotypes in increasing or reducing the accuracy of person perception are far more complex and interesting, and the actual data cast laypeople in a far better light than these sort of caricatured straw arguments would suggest. And so those issues are addressed in detail in the next chapter.

Notes

1. For the statistically inclined, nowhere in this chapter do I compute average correlations by first performing Fisher's r -to- z transformation, then back-transforming the average z to an r , despite the fact that this procedure is common practice. When I averaged correlations, I took the simple, mathematical average (e.g., if one correlation was .5 and another .8, I report the average as .65). I did this for several reasons. First, it is so simple that anyone with a minimal knowledge of statistics can do it for themselves. Second, r to z is necessary to obtain standard errors, which are necessary for testing whether statistically significant differences between correlations exist. In this chapter, however, such tests were not needed and not performed. Third, simple averaging without r -to- z transformation is *statistically conservative*. It produces a systematic tendency to very slightly underestimate the true correlation (e.g., Silver & Dunlap, 1987). Monte Carlo studies have shown that, with sample sizes over 30, as are all such studies reported here, such bias averages well under .01. Nonetheless, if there is any "flaw" in my use of simple averages, it is that such usage biases the results reported in this chapter *against* finding evidence of stereotype accuracy. Thus, please keep in mind that criticizing this method is tantamount to arguing that the data provide even *more* evidence of stereotype accuracy than I conclude. Be careful what you wish for . . .

2. Ryan (1996) also assessed the relation of self-reported familiarity with the out-group to stereotype accuracy. These results showed only modest and inconsistent relationships to stereotype accuracy.

3. For the statistically disinclined, meta-analysis is a technique for combining results from many studies addressing similar topics in order to identify average, or typical, effect sizes. In this case, Swim (1994) took advantage of the prior existence of many meta-analyses of sex differences. For example, many studies have examined sex differences in aggressiveness. The meta-analyses average the effects of those many studies to estimate the overall, or average, difference between men's and women's aggressiveness. Swim then compared people's beliefs about men's and women's aggressiveness to the real difference as indicated by meta-analysis.

18 Stereotypes and Person Perception

CAN JUDGING INDIVIDUALS ON THE BASIS OF STEREOTYPES *INCREASE* ACCURACY?

THE PRIOR THREE chapters (1) concluded that it was unjustified and counterproductive to define stereotypes as inaccurate and (2) reviewed dozens of studies demonstrating moderate to high accuracy in stereotypes. This chapter addresses a different question: Does relying on a stereotype when judging an individual person necessarily reduce the accuracy of that judgment?

Like most of the issues addressed in these chapters on stereotypes, the answers may seem obvious. Is it not true that we should judge individuals entirely on their individual merits? Is it not true that we should *never* allow our stereotypes to influence, bias, or color our judgments regarding individuals? Isn't relying on stereotypes the type of thing that only a racist or sexist would do?

In fact, I do not think the answers to questions like these are obvious at all, at least not if one spends some time and effort to think through these issues and examine the empirical evidence. Let's start with some concrete examples.

Stereotypes and Person Perception: How Should People Judge Individuals?

SHOULD?

"Should" might mean many things. It might mean, "What would be the most moral thing to do?" Or, "What would be the legal thing to do, or the most socially acceptable thing to do,

or the least offensive thing to do?" I do not use it here, however, to mean any of these things. Instead, I use the term "should" here to mean "what would lead people to be most accurate?" It is possible that being as accurate as possible would be considered by some people to be immoral or even illegal (see Chapters 10 and 15). Indeed, a wonderful turn of phrase, "forbidden base-rates," was coined (Tetlock, 2002) to capture the very idea that, sometimes, many people would be outraged by the use of general information about groups to reach judgments that would be as accurate as possible (a "base-rate" is the overall prevalence of some characteristic in a group, usually presented as a percentage; e.g., "0.7% of Americans are in prison" is a base-rate reflecting Americans' likelihood of being in prison). The focus in this chapter is exclusively on accuracy and not on morality or legality.

ON THE USE OF INACCURATE VERSUS ACCURATE STEREOTYPES IN JUDGING INDIVIDUALS

Relying on inaccurate stereotypes will not increase accuracy in judging individuals. This can be readily seen with a nonsocial example. If Fred believes that Anchorage, Alaska, is warmer than New York City, and he relies on that belief for making guesses about where it is going to be warmer, today, tomorrow, the next day, etc., he will be wrong most of the time. Even though he may pick up an occasional hit on the rare days that Anchorage really is warmer than New York, he will be wrong far more often than he is right. He will be wrong far more often than if he obtained a weather report, he would be wrong far more often than if his belief was accurate, and he would be far more wrong than even if he guessed that their temperatures were equal.

Stereotypes are no different. If Elmer believes that professional (American) football players are unusually tiny, and if he relies on that stereotype to guess their sizes, he will usually be very wrong.

Relying on an erroneous generalization to predict a particular case will usually lead one to an erroneous prediction. This point may provide some insight into why using a stereotype for judging a person is so commonly viewed as inherently bad and inaccurate. As discussed in Chapter 15, many laypeople and, indeed, many social scientists *assume* that stereotypes are inherently inaccurate. Given this assumption, it is completely logical to condemn anyone for using stereotypes to judge an individual, because *relying on an inaccurate stereotype will almost always decrease accuracy in judging an individual*.

It may be logical, given the assumption of stereotype inaccuracy, but it is completely wrong, for two reasons: (1) Defining stereotypes as inaccurate is itself unjustified (Chapter 15) and (2) empirically, there is abundant evidence that stereotypes are often accurate (Chapters 16 and, especially, 17). So, blanket condemnations of people for using stereotypes to judge individuals is itself deeply flawed, if it relies on the unjustified assumption that all or most stereotypes are inaccurate (it is *even more* flawed if it does not rely on this assumption, as shall soon be demonstrated).

There is no controversy regarding the unreasonableness of relying on an inaccurate stereotype for judging individuals. Therefore, the remainder of this section focuses on understanding the conditions under which relying on an accurate stereotype will increase or reduce the accuracy of person perception judgments.

THE THREE PRIMARY SITUATIONS

Understanding the role of stereotypes in increasing or reducing accuracy in judgments regarding a person is facilitated by considering three different person perception situations. People will be most accurate doing something different in each of these three situations (because this discussion involves understanding the role of stereotypes in person perception, in all situations, people know the person's group membership). In one situation, people may have access to abundant, clear, relevant *individuating information* about a particular individual. The term "individuating information" refers to information particular to a target person, rather than his or her group memberships. It includes features such as a person's personality, preferences, tastes, attitudes, accomplishments, experiences, competencies, and behaviors. So, when we have abundant individuating information, we really know a lot about the unique characteristics of the person.

In a second situation, we might know something, but not much, about the person. We might, for example, have heard that the person acted assertively once, or we might be selecting people for job interviews exclusively on the basis of their resumes, or we might discover a person's score on a test. In such situations, our individuating information about the person is above zero, but it is also incomplete, missing much information that could be relevant.

In yet a third situation, we might know almost nothing about a person, other than his or her group membership. That is, we have no useful individuating information.

The extent to which people should rely on accurate stereotypes to be as accurate as possible when judging individuals is different in each of these situations. So, each is discussed next, in turn. In each case, I first present an example involving nonsocial perception, in which the issues may, perhaps, be easier to understand and which will certainly be less polluted by political correctness concerns. For the statistically inclined, the logic here follows that of Bayes' theorem, although the presentation here is conceptual and not mathematical (see, e.g., Krueger, 1996; McCauley, Stitt, & Segal, 1980, for explicitly Bayesian analyses of stereotypes).

DEFINITIVE INDIVIDUATING INFORMATION

The first situation involves having vividly clear relevant individuating information about a particular target. I refer to such individuating as "definitive" because it provides a clear, valid, sufficient answer to whatever question one has about a target. For example, when judging academic accomplishments, we might have standardized test scores and class rank and GPA for a college applicant; when judging sales success, we might have 10 years of sales records for a salesperson; and when judging personality, we might have multiple judges' observations of and well-validated personality test scores for a particular individual. When we have this information, how much should we rely on stereotypes?

Alaska and New Jersey. If one discovers from a credible source (say, the Weather Channel) that it is 80 degrees today in much of Alaska, but only 60 in New Jersey, what should one conclude? The answer is obvious. The fact that it is usually colder in Alaska is not relevant. Today, it is warmer in Alaska.

Stereotypes and person perception. Professional basketball players tend to be tall—very tall. It is very rare to find one shorter than 6 feet 4 inches, and if you have ever met anyone that

tall, you know that is very tall. It is, therefore, reasonable to expect professional basketball players to be very tall.

But, every once in a while, a truly short guy makes it into the National Basketball Association. Spud Webb was a starting player in the 1990s, and he was about 5 feet 7 inches. Now that you know his height, should your stereotype of basketball players influence your judgment of his height? Of course not. His height is his height, and his membership in a generally very tall group—NBA players—is completely irrelevant.

In situations where one has abundant, vividly clear, relevant individuating information, the stereotype—its content, accuracy, etc.—becomes completely irrelevant. One should rely entirely on the individuating information.

USEFUL BUT NOT DEFINITIVE INDIVIDUATING INFORMATION

In many other situations, people may have useful information, but it may fall short of being as definitive as in the first situation. Sometimes, one simply does not have much information, or that information is ambiguous, or one is trying to understand or predict something that is fundamentally not capable of being definitively known (and these situations are not mutually exclusive—one may need to make a judgment, decision, or prediction about something that cannot be definitively known and may have only a small bit of ambiguous information on which to do so). Variations on this situation are discussed next.

Small amounts of information. Sometimes, one may simply not have that much information. For example, when we meet a person for the first time, we might have only physical appearance cues (which will usually reveal sex and approximate age, but which may or may not clue us in on race/ethnicity, attractiveness, neatness, concern with fashion, etc.). Or, although we may not be following election for town council closely, we just happen to hear on the radio a 10-second sound bite from a candidate for town council in which she claims that property taxes are too high.

Ambiguous information. Some information is inherently ambiguous—its meaning and interpretation are unclear. Is a shove playful horsing around or assault? Is that a warm, friendly smile or a superior sneer? Is that extreme compliment flattery or sarcasm? There often is no clear answer to questions such as these.

Inferences versus observations. Behavior can be observed directly. Most other aspects of psychology—beliefs, attitudes, motivations, personality, intentions, etc.—are not directly observable. They must be inferred on the basis of behavior. Whereas it is possible to definitively know (most of the time) whether David smiled, without lots of other information, it is not so easy to figure out whether David is “happy.” Whereas it is relatively easy to grade a student’s test, without lots of other information, it is quite hard to know whether that high test score reflects the student’s brilliance or the easiness of the test. There is an inherent ambiguity in going from behavior to inferences about underlying attributes.

Predicting the future versus evaluating the past. The future is inherently ambiguous. It is not possible to know exactly what will happen in the future (history is littered with the inaccurate predictions of the holy [first- and second-century predictions that “Jesus will soon return”], the greedy [“the stock market is going up {or down} this year”], the political [“the Iraqis will greet us with open arms as liberators”], and the superstitious [“because your moon is in Virgo, you will find your lifelong love this week”])).

Nonetheless, we must make predictions about the future all the time. Whenever we select people for admission to college, graduate school, or jobs, we are, essentially, making a prediction that that person is the best for the college, program, or job, or, at minimum, that they are likely to be able to succeed at college, graduate school, or the job reasonably well. Because the future is inherently unknowable, however, we can almost never have enough information to render such predictions definitive.

Although there are some differences among these types of situations (small amounts of information, ambiguous information, inferences, predicting the future), with respect to understanding whether relying on accurate stereotypes increases or reduces person perception accuracy, they share a fundamental, underlying similarity: They all involve making a judgment under uncertainty, that is, in a situation in which the individuating information does not provide a definitive answer to the question. Will relying on an accurate stereotype in such situations increase or reduce accuracy of person perception?

Alaska versus New York. Again, your task is to figure out where it is colder. You get one piece of information about each location. You learn that Jane, a lifelong resident of Anchorage, considers it “cold” today. You also learn that Jan, a lifelong resident of New York, considers it “cold” today.

Note that the “information” that you have is essentially identical regarding the two places. Should you, therefore, predict that they have identical temperatures? Social psychologists routinely condemn people for using stereotypes to judge others in the absence of much clear individuating information (e.g., American Psychological Association, 1991; Darley & Fazio, 1980; Fiske & Neuberg, 1990; Jones, 1986; Jost & Kruglanski, 2002). If you accept this logic and apply it here, you would be compelled to conclude that people could only be “unbiased” (read: “accurate”) if they concluded that the temperatures were the same in both places.

That, however, would be a silly conclusion, because it completely ignores the wealth of information you already bring to bear on the situation:

1. It is usually much colder in Anchorage.
2. “Cold” can mean lots of different things in different contexts.
3. People usually adapt to their conditions, so, if it is usually 40°F in your neighborhood, you would probably judge 20°F as cold; but if it is usually 60°F in your neighborhood, 40°F might be seen as quite cold.

To ignore all this would be foolish. And, most of the time, *ignoring* this will lead you to an inaccurate conclusion about the weather in the two places. Instead, if one relies on one’s accurate knowledge about general differences between Alaska and New York, one is far more likely to be correct than if one ignores that information.

In this case, you probably would be best off relying both on the specific information you received (both residents considered it “cold,” so it is probably colder than usual in both places) and on the valid generalization that Alaska is usually colder than New York. Let’s say “today” is March 1, and the average temperature in Anchorage is usually 30 and in New York it is 45. A sensible and logical prediction, given the information you have, would be that it is about 10 degrees in Anchorage and 25 degrees in New York. You could be wrong, but if your general belief about a 15-degree difference between the cities is correct, and as long as Jane

and Jan are reasonably credible gauges of their local temperature, your estimate that Anchorage is colder than New York is far more likely to be correct than had you estimated that they would be the same.

Stereotypes and person perception. The logic here is identical. Consider stereotypes of peace activists and Al Qaeda members. You hear the same thing about an individual from each group: that they have “attacked” the United States. Should you interpret this to mean that they engaged in identical behaviors? Not likely. The “attack” perpetrated by the peace activist is most likely a verbal “attack” on U.S. war policies; the Al Qaeda attack is probably something far more dangerous. As with temperatures in Alaska and New York, you should use both the valid generalization (in this case, a stereotype) and the individuating information. Knowing that either group “attacked” the United States is very different than knowing that they “praised” the United States. Nonetheless, predicting that an Al Qaeda attack would be identical to a peace group “attack” would be silly. Using your stereotypes will likely aid, not hinder, you in reaching a valid understanding of the nature of the attack.

The same principles hold regardless of whether the stereotypes involve groups one chooses, such as peace activist or Al Qaeda, and about whom it is sometimes socially acceptable to hold stereotypes, or groups one does not choose and/or about whom it is socially unacceptable to hold stereotypes (sex, nationality, race, social class, religion, ethnicity, etc.). If we learn that their friends describe both “Bob” and “Barb” as “tall,” what should we conclude? Should we conclude that they are exactly equal in height? Of course not. Most likely, Bob is tall for a man, and Barb is tall for a woman. Because men are, on average, taller than women, “tall” means different objective heights for men and women (implicit acceptance of these “shifting standards” has been thoroughly demonstrated, e.g., Biernat, 1995).

What about judgments about more socially charged attributes, such as intelligence, motivation, assertiveness, social skill, hostility, etc.? *The same principles apply.* If the stereotype is approximately accurate *and* one only has a small bit of ambiguous information about an individual, using the stereotype as a basis for judging the person will likely enhance accuracy.

Let’s assume that 30% of motorcycle gang members are arrested for criminal behavior at some point in their lives, and .03% of ballerinas are arrested for criminal behavior at some point in their lives. Now, let’s assume, when arriving at a train station, you get a small piece of ambiguous information about each—waiting for the next train at one end of the station is a group of bikers and, at the other end, a group of ballerinas.¹ People are being completely reasonable and rational if, when alone, they avoid the motorcycle gang members and head over toward the ballerinas.

In all of these cases, the stereotype “biases” the subsequent judgments. At least, that is how such influences have nearly always been interpreted in empirical social psychological research on stereotypes (see, e.g., Fiske & Neuberg, 1990; Devine, 1995; Gilbert, 1995; Jones, 1986; see also Chapters 5 and 9). It is probably more appropriate, however, to characterize such phenomena as stereotypes “influencing” or “informing” judgments. Such effects mean that people are appropriately using their knowledge about groups to reach as informed a judgment as possible under difficult and information-poor circumstances. If their knowledge is reasonably accurate, relying on the stereotype will usually increase, rather than decrease, the accuracy of those judgments (see also Jussim, 1991, 2005).

NO INDIVIDUATING INFORMATION

Alaska and New York. If you are given absolutely no information and are asked to predict today's high temperature in Anchorage and New York, what should you do? If you know anything about the climate in the two places, you will predict that it will be warmer in New York. Indeed, to be as accurate as possible, you should predict this *every time you are asked to do so*. Would this mean your beliefs about climate are somehow irrationally and rigidly resistant to change? Would this mean, to paraphrase the American Psychological Association (1991, p. 1064), you "perceive them as all alike or to expect them to be all alike, which they never are"? Of course not. All it means is that you recognize that, when two regions differ and you are asked to predict the day's temperature and are given no other information, it will always be better to guess that the place with the higher average temperature is warmer.

Stereotypes and person perception. If you are given no information other than race and you are asked to predict who has greater yearly income, Leroy Rashid Jefferson, who is African American, or George Spencer Billingsworth III, who is White, what should you do? If you know anything about the income of African Americans and Whites in the United States, you will predict that George is richer. Indeed, you should predict this every time you are asked to make a prediction about the income of an African American and White target about whom you have no other information. Would this mean your beliefs about racial differences in income are somehow irrationally and rigidly resistant to change? Would it mean that you perceive all African Americans as alike? Would it mean that you perceive all White Americans as alike? Of course not. All it means is that you recognize that, when the average income of two racial groups differs and you are asked to predict the income of an individual from those groups and are given no other information, it will always be better to guess that the person from the group with the higher average income has more income.

In such situations, if you always predict the same outcome, you will be wrong sometimes. There are millions of middle class and wealthy African Americans; there are tens of millions of poor Whites. If one African American and one White individual were repeatedly selected at random from the U.S. population, there would be some times when the African American selected was wealthier than the White selected. More often, however, the selected White would be wealthier. Always predicting that the selected White would be wealthier would not be perfectly accurate—but doing so would lead one to be as accurate as possible under the circumstances.

Or, as Kahneman and Tversky (1973, p. 243) put it in their famous article on the psychology of prediction: "When uncertainty is maximal, a fixed value is predicted in all cases." This is what cognitive psychologists and statisticians refer to as normatively appropriate. It means people are doing the best they possibly can—reaching the most accurate predictions possible—under information-poor circumstances. It does not reflect some sort of irrationally rigid bias or error.

WHAT SHOULD PEOPLE DO TO BE ACCURATE? CONCLUSIONS

1. People should not use an inaccurate stereotype to judge a person. This will usually undermine accuracy.

2. People also should not use an accurate stereotype to judge a person when they have abundant, clear, relevant individuating information about that person. This will not improve accuracy, either.
3. When they have little information about a person, or when the information they do have is ambiguous, however, relying on an accurate stereotype will generally increase accuracy in judging a person.

There is, however, a problem. Abundant evidence on overconfidence and naive realism (Ross & Nisbett, 1991) suggests that people often overestimate the accuracy of their own beliefs. Therefore, the fact that a person *believes* his or her stereotype is accurate is not equivalent to the stereotype actually being accurate. Furthermore, people will sometimes, and perhaps often, believe their stereotypes are accurate when they are not.

So, what is a person to do?

Recommendations for a reasonable person. Inaccurate stereotypes do not help person perception accuracy, and many of our beliefs may be less accurate than we think. Therefore, our hypothetical reasonable person should *pay attention to data*, both when developing beliefs about groups and when judging individuals. Regarding groups, as much as possible, one needs to stick to systematic data from credible sources (e.g., Census data, research results, meta-analyses, news reports summarizing scientific research, etc.).

Absent that, and given that we all can't be social scientists poring over Census statistics or research tomes, one has little choice but to reach conclusions about groups from personal experience and other nonscientific forms of "data." Especially given the well-established existence of a slew of systematic errors and biases in the ways people select data, integrate the data, and interpret data (e.g., Kahneman, Slovic, & Tversky, 1982; Nisbett & Ross, 1980), such conclusions should be held gently, with an easy readiness to change them on the basis of more credible sources of information about groups.

When judging an individual, when possible, a reasonable person should get the relevant individuating information. No one is ever fully defined by stereotypes of his or her groups. Individuating information that is clear, valid, abundant, and relevant will almost always be a better basis for judging an individual than are the general characteristics of the groups to which that individual belongs.

Of course, individuating information is not inherently some sort of gold standard, either. First, some types of individuating information, such as that contained in a job resume, are potentially subject to manipulative distortions. Second, just as people overestimate the accuracy of their own stereotypes, they also overestimate their ability to reach valid conclusions from certain types of individuating information, particularly personal interviews (e.g., Ross & Nisbett, 1991). Sometimes, therefore, it might indeed be better to rely on an accurate belief about a group when judging an individual than it might be to rely on deeply flawed individuating information.

Consider politicians running for office. Does one get more information about their likely policy positions from their public statements (individuating information) or from knowing their political party (stereotype about a group)? The answer is not obvious, because *sometimes* the answer is statements (especially when they take controversial or unpopular positions, or positions counter to those of their party), other times party (especially when said politicians attempt to slant their positions to get more votes). But, if party is *ever* more

useful, then we have a situation where the stereotype is more useful than the individuating information.

Sometimes, therefore, individuating information may not be very useful, it may be manipulated or distorted, or it may be only partially and incompletely related to whatever one is trying to judge. Therefore, it behooves our hypothetical reasonable person to become conversant with the types of information that are more versus less useful when judging a person.

But what should a reasonable person do in the absence of clear data, regarding either the group or the person? Personal experience, though subjective and flawed in some ways, is often all we have to go on. And, if one is careful and thoughtful about how one uses personal experience, there is no reason not to use it, too. Has one discovered that 11-year-old boys play soccer more aggressively than 11-year-old girls? If so, then it is reasonable for our reasonable person, who just happens to be a soccer coach, to expect the new 11-year-old boys on the team to be more aggressive than the new 11-year-old girls. Although our coach is being reasonable by holding such expectations, our reasonable coach also needs to understand that these are just predictions. The future is, of course, inherently not knowable with certainty. Therefore, our reasonable coach needs to recognize that those expectations may be wrong for any particular child or group of children. Therefore, Coach Reasonable should be especially interested in and sensitive to new individuating information as it comes in regarding aggressive play. If Rachel is an especially aggressive player, Coach Reasonable should see her that way.

In addition, to the extent that aggressiveness is an important aspect of soccer, Coach Reasonable may want to develop a coaching plan specifically geared toward increasing the aggressiveness of kids who are insufficiently aggressive, expecting (but not knowing for certain) that, disproportionately, those kids will be girls. The last thing Coach Reasonable wants to do (if Reasonable wants his team to be the best it can be) is to create a self-fulfilling prophecy that discourages the development of aggressiveness among the kids Reasonable thinks are already not very aggressive. And, perhaps, by the end of the season, if Coach Reasonable is a good coach, many of the less aggressive kids will have become more aggressive and better soccer players. In this situation, the expectation, held tentatively and flexibly, has helped, not harmed, the players on the team, even the (initially) less aggressive ones.

Recommendations for social science interpretations of the literature on stereotypes and person perception. Much of the social science literature is written with a tone suggesting that people are doing something wrong, irrational, or malicious if they rely on their stereotypes *at all* when judging an individual. Doing so is identified as a source of concern (Darley & Fazio, 1980; Stangor, 1995), it is allegedly unjustified (Fiske, 1998; 2004; Fiske & Neuberg, 1990; Fiske & Taylor, 1991), and it is something to be prevented when possible (Borgida, Rudman, & Manteufel, 1995; Gilbert, 1995).

At least some, and perhaps most, of this predisposition to view using stereotypes as something negative derives generally from the assumptions that they are inaccurate and irrational, foundations of discrimination and prejudice (assumptions that are tenuous at best, e.g., Chapters 15 through 17; Park & Judd, 2005), and, particularly, from the civil rights legislation of the 1960s. That legislation specifically prohibited hiring (and other forms of discrimination) intentionally on the basis of race, sex, religion, and national origin. In 1963, you could hang a sign saying "Manager wanted, Blacks, Jews, and Catholics need not apply," but by 1966, it was illegal to hang such a sign.

This was undoubtedly a good thing—one of the best American domestic political developments of the 20th century. A common clarion call of the 1960s, and later, was that we should not “judge people on the basis of their race” (or any other category). And that is true—we should not judge people on the basis of their race (or any other category). Nothing in this book contests that.

However, over the years, the idea that we should not discriminate has risked morphing into the idea that, but for our different circumstances, we are all fundamentally the same. Such logic is apparent any time researchers or pundits interpret evidence of group differences as reflecting discrimination (e.g., Greenwald & Krieger, 2006; Kang & Banaji, 2006; Sidanius & Pratto, 1999).

Social psychology’s long-standing emphasis on the power of situations over individual behavior (including, but not restricted to, the fundamental attribution error, conformity and obedience research, work on priming and automaticity, etc.—e.g., Ross & Nisbett, 1991, but see any undergraduate text on social psychology) has done more than its part to provide support to this view (and see Krueger & Funder, 2004, for a cogent critique of social psychology’s excessive situationism). We all want a good life and our families to be safe and prosperous. We all want freedom and dignity. And so on. But for our situations, the story goes, we would all be the same, or at least nearly so. And, we have unequal outcomes largely or entirely because discrimination systematically disadvantages (or advantages) some of us.

Regardless of the benevolent intentions of those promoting such a view, the idea that we are all so much the same that our differences hardly matter is not justified. For example, the entire point of the “diversity” rationale for preferential selection of particular classes of individuals for admissions, hiring, promotion, etc., is that people from diverse backgrounds bring in different cultural experiences and ways of thinking that enhance everyone’s experience (whether educational or occupational). Keeping in mind that diversity is usually “measured” by a person’s demographic characteristics (especially race and ethnicity, sometimes sex, occasionally other demographics), this notion is deliciously ironic. It is, essentially, an argument for stereotyping! It says we need not rely on individuating information to assess “diversity”—no need to ask written or interview questions about people’s experiences, backgrounds, knowledge, points of view, etc. No need to create psychological tests assessing diversity of background. Instead, we can infer their “diversity” of background, experience, knowledge, etc., from their race, religion, or other social category membership. Those advocating this type of diversity are implicitly advocating stereotyping writ large.

Ironies aside, however, the diversity argument only makes sense by assuming that people are not all the same. They differ in ways so fundamental and important that we should make extra effort and go to extra expense to ensure our schools and workplaces are populated with people from diverse racial, ethnic, and cultural backgrounds. Fine. But one cannot then turn around and make the argument that we are all essentially the same.

If we are all the same, it is indeed incorrect and dysfunctional to use a stereotype at all, ever, to make inferences or predictions about a particular individual. But the second one acknowledges that we are not all the same, and that it is possible for a stereotype to accurately capture bona fide group differences, this view collapses. Instead, one is compelled to also acknowledge that *failing* to use that stereotype to judge a person, except when we have abundant and vividly clear individuating information, is most often going to lead to a *less* accurate judgment than is using that stereotype.

So, after all this discussion, I can present my recommendations for social scientists:

1. Stop claiming or implying that any use of a stereotype somehow constitutes something bad, unjustified, inaccurate, and immoral.
2. If you really want to claim that reliance on a stereotype is bad in a particular instance, you must first do either one of two things. First, you must provide clear, empirical evidence that people actually hold a stereotype (i.e., subscribe to a certain belief about a group), and then show that belief to be factually invalid. In that case, any use of it will be unjustified. Those who claim that accuracy cannot or should not be studied (e.g., Fiske, 1998, 2004; Stangor, 1995) have, however, removed from themselves the possibility of providing such evidence (unless, of course, they change their views regarding the possibility of assessing [in]accuracy). Second, you could attempt to justify a claim that the individuating information you are studying is 100% diagnostic of whatever is being judged. Such situations are probably rare, but they also do probably exist. If so, then relying on a stereotype, even an accurate one, cannot increase accuracy in that judgment.
3. If you have a reasonably good understanding of the basic principles of logic and statistics (e.g., Bayes theorem), and if you have read Kahneman and Tversky's (1973) classic work identifying people's failure to use base-rates, you must fully articulate the (apparently, very different set of) logical and statistical principles that you are using when you condemn laypeople for using base-rates (their beliefs about the general characteristics of a group) to make inferences about an individual in the absence of perfectly clear and relevant diagnostic individuating information.
4. If you really desire to reach the conclusion that people's judgments are *factually wrong or unjustified*, perform an accuracy study. Or, at minimum, review studies that actually assess accuracy, rather than merely reviewing studies that assess processes that researchers presume cause inaccuracy without actually testing for accuracy. Recognize and acknowledge that an effect of a stereotype on judgments of a person cannot be known to undermine accuracy absent an assessment of accuracy.

What Do People Do? The Data on Accuracy, Reasonableness, Error, and Bias in People's Use of Stereotypes and Individuating Information to Judge Individuals

Thus far, this chapter has focused primarily on understanding what people should do—how they should use stereotypes and individuating information—to be as accurate, rational, and reasonable as possible. This analysis was necessary for several reasons. First, as discussed in Chapters 10 through 12, if one wants to claim that people are doing something wrong, invalid, irrational, or inaccurate, one needs to first articulate what would be right, valid, rational, or accurate. Only after doing so can one reach any conclusions about how close people come to this ideal.

Second, although models of rationality and accuracy are common in the decision-making literature, they are almost entirely absent from the stereotyping literature (see Chapters 10 through 12). Absent such a standard, it is all too easy and, indeed, too common for social

scientists to condemn people for error and bias *no matter what they do* (again, see Chapters 10 through 12). Third, much like accuracy issues more generally, the absence of many prior clear statements regarding what people are supposed to do with stereotypes has led to the rise of many myths and misconceptions about stereotype use, nearly all of which overemphasize, sometimes in an extreme manner, the inappropriateness of relying on a stereotype for judging an individual (any perspective that states or implies that relying on a stereotype necessarily reduces accuracy—e.g., American Psychological Association, 1991; Aronson, 1999; Fiske & Neuberg, 1990; Stangor, 1995—is clearly overstating the case against stereotype use).

So, the first part of this chapter described conditions under which it is more versus less reasonable and appropriate to rely on stereotypes when judging individuals. Of course, this tells us nothing about what people actually do. Perhaps they massively and overwhelmingly deviate even from this more even-handed view of how and when to use stereotypes. The next section, therefore, addresses the data on what people actually do.

THE EARLY DATA

Before about 1980, understanding the role of stereotypes in person perception was not a big area of research in social psychology. Nonetheless, the few studies that were capable of addressing this issue (whether or not they were framed this way) consistently provided evidence of lots of reasonableness along with some stereotype bias. The LaPiere (1934) Chinese couple study (see Chapter 2) showed that, despite claiming they would not provide service to Chinese people, the service personnel at about 200 hotels, campgrounds, and restaurants readily provided polite and pleasant service nearly every time when faced with an actual Chinese couple. One common way this inconsistency between stated policy (“no service to Chinese”) and the courteous service provided to an actual Chinese couple was explained went something like this: This couple was well-dressed, well-spoken, and not at all what people expected when thinking about Chinese people.

This analysis, then, assumes that the service personnel judged the couple in a completely reasonable, rational, and appropriate manner. The service personnel, apparently, readily jettisoned their stereotypes (and even their prejudices) regarding Chinese people and provided service on the basis of their personal characteristics rather than their race. That is, they judged the couple based on the available individuating information, not stereotypes, exactly as nearly every theoretical perspective says they should.

An active line of research in the 1960s involved the “belief similarity model of prejudice,” which argued that Whites disliked African Americans mainly because Whites assumed African Americans held different beliefs and attitudes. It predicted, and consistently found, that Whites, even Southern Whites, evaluated African Americans who held attitudes similar to their own more positively than they evaluated Whites holding attitudes that differed from their own, and about as positively as those of Whites holding attitudes similar to their own (e.g., Rokeach & Mezei, 1966). Especially in the 1960s American South, that was one powerful individuating information effect.

In the 1950s, Clarke and Campbell (1955) examined White and African American elementary school students’ predictions regarding the African American students. Overall, they found modest evidence of bias and substantial evidence of accuracy (this study is described in more detail later in this chapter).

This leads one to wonder: Why did none of this data figure into those reaching the conclusion that, “once categorized, they are all judged as the same, which, of course, they never are”? Indeed, one might think that, because of studies like these (Clarke & Campbell, 1955; LaPiere, 1936; Rokeach & Mezei, 1966), by the 1970s, the idea that people primarily rely on individuating information when judging individuals would have been well-known, because the data were already well-established and reasonably well-replicated across decades, contexts, and types of groups. Nothing, however, could have been further from the truth.

THE SAGA OF LOCKSLEY

The research. Locksley, Borgida, Brekke, and Hepburn (1980) performed the first studies explicitly framed as addressing the role of stereotypes and individuating information in person perception. In the introduction to the paper, they indicate that they expected to find strong evidence of sex stereotypes biasing people’s perceptions of an individual man and woman. Despite the data produced by Campbell, Rokeach, La Piere, and their colleagues, nearly all of the scholarship up to that time emphasized the inaccurate and irrational nature of stereotypes and their power to distort judgments (see Chapter 15).

Given the social science discourse about stereotypes up to that time, that Locksley et al. expected to find powerful sex stereotype effects was completely plausible. But it is not what they found.

First, they assessed sex stereotypes regarding assertiveness. People believed that, in general, men were more assertive than women. Second, they asked perceivers to evaluate an individual man or woman under one of three conditions: (1) no individuating information, just the name (which indicated sex); (2) useless individuating information (the target got a haircut); or (3) a single instance of assertive behavior (interrupting a student dominating class discussion, i.e., an assertive behavior). Perceivers then rated the assertiveness of the target.

Results were quite clear; without individuating information or with useless individuating information, there was a clear effect of the stereotype: The perceivers rated the man as more assertive than the woman. With useful individuating information, there was no stereotype effect: The perceivers rated the man and woman as equally assertive. Locksley (Locksley, Hepburn, & Ortiz, 1982) then followed this up with another set of studies, this time of stereotypes regarding “day people” versus “night people,” and found essentially the same results. Clear, relevant individuating information eliminated stereotyping.

The controversy. The benevolent view of what happened is this: The research community promptly began performing follow-up studies in an attempt to identify the conditions under which individuating information does and does not eliminate stereotyping. And, undoubtedly, many researchers were deeply and sincerely interested in discovering those conditions.

Another, less benevolent, view is that many researchers promptly went on a quest to limit the “damage” (to the traditional view of stereotypes as being powerful) by performing studies purporting to show the extraordinary difficulty of eliminating stereotype effects (and, when it was not really very difficult, interpreting the research in such a manner as to imply such difficulty). Many narrative reviews emphasized the unusual difficulty of finding individuating information eliminating stereotype effects (e.g., Borgida et al., 1995; Fiske & Neuberg, 1990; Fiske & Taylor, 1984, 1991; Hamilton, Sherman, & Ruvalo, 1990; Jones, 1986; Neuberg, 1994). Fiske and Neuberg (1990) were at the forefront of this effort and, in

their classic review, argued that stereotypes were the default basis for person perception. According to their model, only when the following supposedly extraordinarily difficult-to-obtain set of conditions occurred would people jettison their stereotypes: (1) People had to be motivated to pay attention to individuating information, (2) they needed the (easily expendable) cognitive resources to pay attention to the individuating information, and (3) the individuating information had to overlap completely with the judgment (e.g., if perceivers had IQ information and were asked to rate a target's IQ, then would people jettison stereotypes; otherwise, they would not). It quickly became "common knowledge" that stereotypes were the default basis of person perception and belief in their extraordinary power to influence person perception remained safely intact (e.g., American Psychological Association, 1991; Borgida et al., 1995; Devine, 1995; Jones, 1990; Rahn, 1993).

This is not some sort of "straw" interpretation. The following quote from a famous and influential paper by some famous and influential social psychologists essentially innocently reflects this common and widespread interpretation of the prevailing theoretical perspectives emphasizing how extraordinarily difficult it supposedly is to eliminate people's reliance on stereotypes:

"As Fiske (1989, p. 253) described, 'stereotypers categorize because it requires too much mental effort to individuate.' . . . A characteristic feature of cognitive models of impression formation is the priority they accord to category-based processes in person perception (Brewer, 1988; Fiske & Neuberg, 1990). Perceivers seem at best reluctant, and at worst incapable, of individuating others unless a series of critical cognitive and motivational criteria (e.g., spare attentional resources, self-involvement, outcome dependency, and accountability) have been satisfied."

(Macrae, Milne, & Bodenhausen, 1994, p. 44)

Of course, this is a conditional statement: "People are reluctant/incapable to individuate *unless*" a series of conditions are met. Maybe those conditions are met most of the time in real life. But that is not the tone of such interpretations or the articles on which they are based. The connotations are all about the alleged extraordinary difficulty of getting people to individuate. It is, perhaps, worth noting that in none of Locksley's studies did she do anything to ratchet up people's attentional resources, self-involvement, etc. She just had people read information about targets and judge them.

Locksley is no longer a social psychological researcher. It is likely that the controversy these studies generated also came with some hostility, and that did not do much to help keep her in the field. The rumor mill had it that she was disgusted with academics, decided she did not need the aggravation and, in fact, went on to a very nice life outside of academics. I hope this is all true—but what a loss for social psychology.

So, let's get back to data. What do they say? To date, hundreds of studies of the role of stereotypes in person perception have been performed. Do they show stereotype effects to be powerful and difficult to eliminate? Or do they mostly back up Locksley, showing that people rely primarily on individuating information when judging others?

First, I review the broad and general patterns. This is a great catch-all, because, as shall be seen, the broad and general pattern shows stereotype effects to be even *weaker* than those found in the particular studies that I will review. Second, I review a small number of

particular studies, because they are especially relevant to understanding processes of stereotype use or the accuracy produced by stereotype use or disuse.

GENERAL PATTERNS

The meta-analyses. About 300 studies have addressed the role of stereotypes in person perception. We have already seen this data, in the bottom half of Table 6-1 (in Chapter 6). When considering them all together, they show that, on average, stereotypes have only a very small influence on person perception. This will probably be a shock to many people who have spent their careers touting the evils of stereotypes or who have relentlessly emphasized the supposedly extraordinary difficulties involved in getting people to jettison their stereotypes when judging individuals, or to consumers of such perspectives who have innocently accepted those conclusions at face value.

Nonetheless, the overall effect of stereotypes on person perception, across nearly all the studies of stereotyping that have been performed, averages to a correlation (between target group label and perceiver judgment of an individual) of .10. Furthermore, even this .10 effect is probably an overestimate, because the correlation of the bias effect with the number of studies included in each meta-analysis shown in Table 6-1 is $-.39$. The more studies, the *smaller* the average biasing effect of stereotypes, which again suggests bias in favor of publishing studies demonstrating bias (it suggests that when there are few studies in some domain, they are more likely to provide evidence of bias; as more and more studies are conducted on some topic, the data slowly creep in showing how much smaller bias actually is than first believed).

Regardless, this overall effect is small by any reasonable standard, which can be seen in several different ways. First, it can be interpreted to mean that, overall, stereotypes substantially affected 5% of the judgments in those 300 studies (Rosenthal, 1991). This, of course, means the same thing as concluding that stereotypes did not substantially affect 95% of the judgments. Second, it means that, on average, stereotypes lead to about two tenths of 1 standard deviation difference in how people view targets. Such an effect is “small” by Cohen’s (1988) system of classifying effect sizes.

Third, an effect size of .10 places stereotype effects among the smallest effects obtained by social psychologists (Richard, Bond, & Stokes-Zoota, 2003). Fourth, this effect size is so low that it means that (1) a great many studies find no significant stereotype effects at all and (2) there are nearly as many studies finding reversals of stereotype effects (e.g., people rating an individual woman as more assertive than an individual man, a Latino defendant as more innocent than a White defendant, etc.), as there are studies finding stereotype-consistent evaluations and judgments. Not quite as many, but close.

Let’s make this meaningful. The standard deviation on SATs and GREs (prior to 2010, when the scoring system changed on the GREs) was 100 points, so two tenths of a standard deviation is 20 points. This .10 effect, therefore, means that, even when their true scores are identical, people will, on average, perceive members of THIS GROUP as having scores 20 points higher than THAT GROUP. Twenty points? There are probably some situations in which this tiny difference in SAT scores is judged to be meaningful, but those situations are rare.

In this context, claims that stereotypes exert some sort of extraordinary influence on person perception, and those that emphasize difficulty in limiting stereotype effects, do not

seem to rest on much scientific terra firma. Instead, it seems that stereotype effects on person perception are, in general, weak and easily eliminated. So, does that end the discussion? Not at all.

OBJECTIONS, LIMITATIONS, AND ALTERNATIVE EXPLANATIONS THAT SEEK TO MAINTAIN A BELIEF IN POWERFUL STEREOTYPE EFFECTS AND PERVASIVE IRRATIONALITY

Accumulation of small biases? Social psychology has long been enamored of the idea that small biases accumulate to produce large effects (see Chapter 14). To some extent, this is a self-serving idea, in that it “justifies” the value of research demonstrating only small biases. Nonetheless, if small biases accumulate over time to produce big differences, then, in fact, small biases can indeed become quite important.

Do they? As already discussed (Chapter 14), small self-fulfilling prophecies do not generally accumulate over time. In the classroom, such effects dissipate. And although effects do likely accumulate across perceivers, current evidence is that such effects are typically quite modest.

What about stereotype biases? There is no evidence that directly bears on this issue. A narrative argument for the idea that stereotype effects are likely to accumulate has been made by Claire and Fiske (1998, summarized in Chapter 14). It is, perhaps, worth pointing out that, although Claire and Fiske (1998) cited many studies, none provided any direct evidence of stereotypes producing self-fulfilling prophecies that accumulate (see Chapter 14 for a comprehensive review of the studies that attempted to do so). The closest one can get to a citation to evidence of accumulation is a *simulation* study (Martel, Lane, & Emrich, 1996). In this simulation, Martel et al. (1996) concluded that if women are subject to a very modest amount of discrimination at every level of employment (interview, entry-level hiring, promotion, etc.), such bias will accumulate to produce dramatic disparities in men’s versus women’s advancement. Given their assumptions, this is most definitely true.

There is, however, a problem here. A simulation has no data, at least not in the sense of observations of real people. Instead, it makes certain starting assumptions, and then makes further assumptions about outcomes or processes, and then shows what would happen if its assumptions and guesstimates are correct. A simulation, therefore, is only as good as its assumptions. So let’s examine those assumptions more closely.

Any claim that has the following structure is logically true:

1. Here is a small effect.
2. If it happens repeatedly, in the absence of countervailing effects, it will become a large effect.

Let’s do our own mini-simulation right now. Let us assume that working out with a particular set of weights increases your muscle strength one tenth of 1% after each workout, one works out three times per week, and one starts out being able to bench press 100 pounds. After a year, according to this simulation, one will be able bench press 117 pounds. Not too impressive, right? But after 5 years one will supposedly be able to bench press almost 500 pounds. “Well,” you say, “if you really worked out that consistently for 5 years, maybe

you could.” Fine. After 10 years, you would be bench pressing over 2,000 pounds. Sorry, folks, that is just not happening. Why not? Because the assumption of a one tenth of 1% increase is not correct, at least not indefinitely. The body has limits. Working out with the same weights will not produce a constant increase in capacity.

So, what should we make of the Martel et al. (1996) simulation? As long as one interprets it very narrowly, it is fine. It is absolutely true that *if* women are subject to the amount of bias they assume at *every* level of their careers, and *if* nothing ever intervenes to counter such influence (no laws, no policies, no benevolent bosses, no hardnosed but merely self-interested bosses who could not care less about justice but just want the best person for the job), then economic differences of about the magnitude “found” in their simulation should occur. It does not, however, provide a shred of evidence that any of that actually does occur. Indeed, bias in the real world could be smaller than found in their simulation, or it could be larger. Absent data, it is impossible to know. As such, although it answers the question, “What would happen if small sex bias accumulated throughout women’s careers?” it constitutes no basis at all for assuming that stereotype biases actually do accumulate.

A narrative review suggesting that stereotype biases are unlikely to generally accumulate much. Inasmuch as there are no hard data on the issue, my discussion is just as speculative as that of Claire and Fiske (1998) and as speculative as is the simulation of Martel et al. (1996). Nonetheless, there may be some value in at least considering how the evidence that we do have might suggest such accumulation is not likely to be large.

First, all the arguments against self-fulfilling prophecies accumulating much (see Chapter 14) also apply here. Second, the mountain of research on stereotypes itself provides evidence strongly suggesting such accumulation is unlikely. Although people do rely on stereotypes when they have little other information about a target, when they have individuating information, they use it, and they use it big time. The effect of individuating information on judgments is one of the largest effects in all of social psychology (discussed later in this chapter). People judge others on their merits, at least for the most part, at least in the hundreds of studies performed by psychologists that have addressed this issue.

In real life, whether it is in school, on the job, or among casual acquaintances, we often have ample, repeated, even abundant opportunities to obtain the individuating information most useful for making a judgment. Let’s say a teacher wrongly assumes boys are better than girls at math. Marie, however, is a math whiz. The teacher, if she is at all like the participants in most social psychological studies of stereotypes, will rely heavily on Marie’s actual performance in math, rather than sex stereotypes, when evaluating her. And, so, without denying the teacher’s potential for some degree of stereotype bias, Marie’s brilliance will eventually shine through (see, e.g., Jussim, Eccles, & Madon, 1996; Madon et al., 1998, for empirical, scientific, real-world examples of just such processes; see also the individual stories in Chapters 7, 9, and 14).

But there is even better and far broader evidence that this process—of stereotype biases getting bigger and bigger over time—has to be greatly overstated. Specifically, nearly every ethnic group currently living in the United States has been subject to negative stereotyping at some point or another. Jews? Supposedly “genetically inferior.” Chinese? Coolies fit for little more than slavelike labor on the railroads. Irish? Need not apply. If the stereotype bias story was true writ large, then none of these (or many other) once stigmatized groups could have possibly dug themselves out of the slums and tenements that the first generations

lived in. The biasing effects of others' stereotypes, ala the Martel et al. simulation, would have created progressively more and more disadvantage for these groups. But it did not happen that way. Why rely on purely hypothetical simulations when the real world provides ample testimony against the idea that stereotype biasing effects necessarily, typically, relentlessly accumulate?

Bias against bias? One of the oddest objections to these data I have ever heard (literally "heard"; it was at a conference, and the objection was raised by a very famous and prestigious social psychologist, and these comments were greeted with a wave of nods of approval by the crowd, which also included some famous and prestigious social psychologists, as if the point was well taken) was that the many studies in the meta-analyses showing weak stereotype effects were biased *against* finding stereotype effects, because social psychologists have long been on a quest to identify conditions that could eliminate stereotyping. Therefore, this argument went, the general literature greatly overrepresented such conditions, with the effect of artificially underestimating the power of stereotypes in the real world.

This is a logically tight argument. To paraphrase Fiske and Neuberg (1990), social psychologists are no fools. So, why do I find this odd? Because its core argument is that *social psychological research on stereotypes has been biased against finding bias!* Social psychologists biased *against* finding or extolling the power of bias? If so, such a state of affairs would constitute a startling reversal. The dominant social cognitive perspective within social psychology has, for decades, been little more than the study of bias (see Chapter 1 or the quotes in Chapters 4 and 5)! Accuracy was dismissed as unimportant for decades (see Chapter 10). Whole books have been written about bias, and bias is a central theme in many undergraduate texts (see Chapters 1, 4, 5, and 10). Perspectives emphasizing accuracy exist but constitute a tiny minority of social psychological scholarship. Social psychologists have defined stereotypes as inaccurate and emphasized their inaccuracy for decades (see Chapter 15). In this context, to suggest that social psychologists have a general bias *against* finding stereotype bias is exceedingly odd. In fact, I think the evidence shows quite the opposite: If social psychologists have any bias, it is in favor of finding and emphasizing bias, and to suggest that their research has been biased against finding bias runs counter to decades of social psychological scholarship.

In fact, perhaps the two most startling things about the meta-analyses showing weak stereotype effects and large individuating information effects are that (1) they were found despite the very large predisposition and preference many psychologists have for finding bias, which strongly suggests that such patterns deserve to be considered unusually credible, and (2) many psychologists remain willing to blithely ignore or dismiss the accumulated data from hundreds of studies and continue to happily tout the power of stereotype biases. As Winston Churchill once said in an entirely different context, "He occasionally stumbled upon the truth, but hastily picked himself up and hurried on as if nothing had happened" (Winston Churchill Leadership, 2008).

Conditions under which. It is true, however, that some conditions under which stereotypes have greater or smaller effects are well-established. Indeed, as the earlier section of this chapter demonstrated, even if people wanted to be completely accurate, there would be some conditions under which stereotypes would have some influence on perceptions. In this spirit, then, let's examine some of those conditions, with respect to two questions: How much do people rely on individuating information? and How much do people rely on stereotypes?

CLEAR, ABUNDANT, INDIVIDUATING INFORMATION

Sometimes, people have abundant, clear, relevant individuating information. For example, they may receive information about a target who engages in some sort of assertive or aggressive behavior (e.g., interrupting a classmate, yelling at a spouse) and then be asked to rate the target's assertiveness or aggressiveness. Or, they may receive information about students' performance on tests and assignments for a class and then evaluate those students' academic achievement.

So, how much do people rely on relevant and useful individuating information? A great deal. The effects of the assertiveness of the targets' behavior in the Locksley et al. (1980) studies were consistently around $r = .5$. Even stronger effects of clear, relevant individuating information have been found in most other studies (see also Chapter 9), which is why Kunda and Thagard's (1996) meta-analysis of dozens of studies of stereotypes and person perception produced an overall effect size of about $r = .7$ for individuating information. Like stereotype accuracy effect sizes more generally, these are among the *largest* effects in all of social psychology (see Table 19-1).

This is worth pausing over and contemplating for a minute. The .10 average stereotype effect is *one of the smallest in social psychology*. The .7 average individuating information effect is *one of the largest*. And yet, there has been a broad consensus in the social sciences that getting people to ignore their stereotypes when judging individuals is extraordinarily difficult and, even worse, that "once people categorize others, they judge those others as being all alike." There is, to put it mildly, a clear mismatch here between the data and the narrative conclusions. This mismatch constitutes yet another, and perhaps the single most, extraordinary testament of social psychology's bias in favor of bias. Nonetheless, let's return to the research addressing conditions under which people are more or less likely to rely on stereotypes.

How much do people rely on stereotypes when they have clear, relevant individuating information? Somewhere between not at all and hardly at all. In the Locksley et al. (1980, 1982) studies of sex stereotypes and stereotypes of day and night people, not at all. Similar patterns have been found in many other studies, both experimental (e.g., Baron, Albright, & Malloy, 1995; Krueger & Rothbart, 1988) and naturalistic (e.g., Jussim et al., 1996; Madon et al., 1998).

Occasionally, however, even in the presence of clear and abundant individuating information, small stereotype effects emerge. For example, even though teachers had ample access to students' performance in class and on standardized tests, teachers' sex stereotypes still had a small biasing effect (of about .10) on their judgments of boys' and girls' math performance (Jussim et al., 1996; Madon et al., 1998). A similar pattern of small bias in the presence of clear individuating information was found for the extent to which children's racial stereotypes bias their perceptions of one another's grades (Clarke & Campbell, 1955). Exactly why these very small stereotype effects persisted even in the face of clear individuating information is unclear, and a question that must be left for future research.

It is, perhaps, worth noting though, that in all three studies, even though stereotypes did slightly bias judgments, the effects of individuating information was (typically) much larger (.4 to .7). This is yet another demonstration of several of the main themes of this book: (1) Biases and accuracy can and often do occur simultaneously right alongside one another,

(2) bias is generally small compared to accuracy, and (3) people are not perfectly rational and unbiased, but they are often pretty damn good.

Returning to “conditions under which,” the bottom line is that, when people have the option of using clear, abundant, relevant individuating information or stereotypes to judge a particular person, they usually rely on that individuating information very heavily. Usually, they do so to the exclusion of stereotypes; occasionally, stereotypes will still exert a small biasing effect on judgments even in the presence of clear individuating information. Whether such effects increase or reduce their accuracy will be discussed later in this chapter.

AMBIGUOUS AND SMALL AMOUNTS OF INDIVIDUATING INFORMATION

Sometimes, individuating information is ambiguous—its meaning or interpretation is unclear. For example, people might receive a work sample or test score, without any standards or norms against which to evaluate its quality. Or they might receive a court case transcript, in which there is both incriminating and exonerating evidence regarding the defendant. Other times, people may have some, but not much, individuating information. Whereas the information itself may be clear (e.g., Bob hit a 90-mile-per-hour tennis serve), how much it actually tells us may be quite limited (we still do not know much about how good a tennis player Bob is; indeed, in this situation, we do not even know if his serve landed in or if he faulted).

How large are individuating information effects in such situations? It is hard to reach broad conclusions about this because most of the research using limited or ambiguous information has held it constant (e.g., Cohen, 1981; Darley & Gross, 1983; Goldberg, 1968). Because it is not possible to test the effects of something held constant, most of the research provided no information about the effect size of ambiguous individuating information on person perception.

One of the few studies that did test for effects of such information, however, showed that even small amounts of individuating information could be quite powerful (Krueger & Rothbart, 1988, Study One). Perceivers’ ratings of targets’ aggressiveness systematically and significantly increased as the target’s behavior became more aggressive (getting a haircut vs. yelling at a spouse vs. hitting someone).

How large are stereotype effects in the presence of ambiguous or small amounts of individuating information? In the presence of a single piece of individuating information reflecting degree of aggressiveness, sex stereotype effects were still quite substantial (Krueger & Rothbart, 1998, Study One). Perceivers believed male targets were more aggressive than female targets, even when the individuating information was identical. (Unfortunately, Krueger and Rothbart [1988] did not provide the information necessary to determine the effect sizes for the individuating information or stereotype effects.)

NO USEFUL OR RELEVANT INDIVIDUATING INFORMATION

Obviously, when people have no individuating information, it makes no sense to even raise the question of how much that information influences judgments. What about when people have individuating information that is not relevant to the judgment (e.g., whether or not the target got a haircut, when trying to judge intelligence or assertiveness)? It might make sense

to evaluate how much people use that useless information, but to do so turns the logic of extolling the virtues of relying on individuating information and ignoring stereotypes on its head. Specifically, if it were found that people did use useless individuating information, it would mean that using individuating information, rather than reflecting reasonableness and rationality, would reflect unreasonableness and irrationality. Indeed, perceivers would be far more reasonable and rational to use an accurate stereotype to judge an individual in this situation and completely ignore the (useless) individuating information.

Researchers have indeed investigated whether useless individuating information reduces or eliminates stereotyping (Hilton & Fein, 1989; Nisbett, Zukier, & Lemley, 1981). Even this research, however, has not examined the effects of different types of useless information. Therefore, it is impossible to know how much people relied on useless information in these studies. So, even though the issue of how much people rely on useless information is potentially interesting, it is one on which we currently do not have much data.

There is, however, a question about which we have quite a lot of data: How large are stereotype effects on person perception when people have little or no useful information? Despite their reputation, even in the absence of much or any individuating information, stereotype effects are typically fairly modest and occasionally nonexistent. Indeed, the early research on this issue seemed to yield contradictory results, sometimes showing that even apparently nondiagnostic (i.e., useless) individuating information can eliminate stereotyping (Nisbett et al., 1981), and at other times showing that nondiagnostic individuating information did not eliminate stereotyping (Locksley et al., 1980). This early controversy was largely resolved by research showing that some types of individuating information (termed “pseudorelevant”), such as intelligence, are often assumed to be so useful for so many types of judgments that, even if it is not specifically diagnostic of a specific judgment, people often use it to some degree (Hilton & Fein, 1989). Pseudorelevant individuating information often eliminates stereotyping, although completely irrelevant information does not.

So what is the main conclusion supported by existing research regarding stereotype use in the absence of useful individuating information? In Kunda and Thagard’s (1996) meta-analysis, stereotype effects averaged $r = .27$ when perceivers had no individuating information at all. So, in general, in the absence of individuating information and in the presence of individuating information that people consider to be useless, people do indeed rely on stereotypes.

Many narrative reviews seem to assume that *any* influence of a stereotype on judgments reduces accuracy. This, however, is not something that can be answered by assumption, and later in this chapter, I explicitly review the small number of studies that have directly examined whether relying on a stereotype increases or reduces accuracy in judging individuals.

CONCLUSIONS REGARDING STEREOTYPES AND INDIVIDUATING INFORMATION: THE STEREOTYPE RATIONALITY HYPOTHESIS

These broad patterns of results are broadly consistent with what I have come to think of as the Stereotype Rationality Hypothesis. According to the analysis presented earlier in this chapter regarding when people should and should not use stereotypes, it is rational and reasonable to use stereotypes in the complete absence of individuating information, when the individuating information is perceived to be useless, and when individuating information is

either scarce or ambiguous. It is also rational and reasonable to jettison stereotypes and rely on the individuating information when that information is clear, credible, relevant, and abundant. This pattern, it seems, closely corresponds to how people actually use their stereotypes—not perfectly (e.g., there are sometimes small stereotype effects even when individuating information is relevant, clear, and abundant), but pretty closely.

In terms of process, people seem to use their stereotypes both gingerly and reasonably. Based on the dramatically larger (on average) effect size of individuating information over stereotypes, rather than stereotypes being some extraordinarily difficult to over-ride automatic default, people seem to strongly prefer judging others on the basis of individuating information. When both stereotypes and individuating information are available, despite claims to the contrary (e.g., Borgida et al., 1995; Fiske & Neuberg, 1990), individuating information appears to be the primary basis for person perception. When individuating information is relevant, people generally use it far more than stereotypes.

Instead, it seems people rely on stereotypes only hesitantly and reluctantly. Only when they have no individuating information or when the individuating information they do have is irrelevant or ambiguous do they use stereotypes to any substantial extent. Stereotypes, apparently, generally function not as a first option but, instead, as a best guess of last resort when there is little else to go on.

This analysis does not preclude the possibility that, chronologically, people may sometimes receive stereotype information before individuating information. In initial face-to-face interactions with strangers, one receives a wealth of demographic/stereotype information instantly (age, sex, possibly race and ethnicity, etc.). That people may rely on this information automatically and without thinking is also well-established (Devine, 1989; Fiske & Neuberg, 1990; Macrae, Stangor, & Hewstone, 1996). If the “default” proponents retreat to this position—that stereotypes are the default in the sense that stereotype information is sometimes received first and relied upon automatically—they would be well-justified.

That, however, is a far cry from claiming that stereotypes are generally tortuously difficult to dislodge or that they constitute common, pervasive, and powerful influences on person perception—positions that are not remotely supported by the data. After adding in the many situations where stereotypes and individuating information are received simultaneously or in which individuating information is received first (e.g., evaluating job applicants on the basis of resumes, college applicants on the basis of transcripts)—situations frequently studied in the empirical literature reviewed in this chapter—it is vividly clear that individuating information, not stereotypes, is the primary basis for person perception.

THE FEW STUDIES OF STEREOTYPES AND PERSON PERCEPTION THAT ACTUALLY ASSESSED ACCURACY

As discussed in Chapter 10, rationality and accuracy are not the same thing. Any given judgment may be arrived at rationally and be wrong or arrived at irrationally and be right. Therefore, showing that people generally apply their stereotypes fairly rationally, although it is good news, does not directly tell us very much about the degree of accuracy of their judgments. Only research that actually assesses accuracy can inform us of the extent to which people's perceptions of groups and group differences end up accurate or inaccurate, and whether relying on stereotypes increased or reduced accuracy in judging an individual.

Unfortunately, however, the widespread assumption of stereotype inaccuracy was a major obstacle to seriously considering the question of whether relying on a stereotype increases or reduces person perception accuracy (see Chapter 15), as was the 30-year period in which studying accuracy was verboten (see Chapter 10). Thus, only a very small number of studies have addressed these issues.

Next, therefore, I review those studies that provided data capable of addressing this issue (this includes several studies that were not framed by their authors as addressing this issue but that, nonetheless, provided data that does address it). The key questions are how accurately did people perceive individuals and groups and whether relying on stereotypes increased or reduced the accuracy with which people perceived group or individual differences.

Clarke and Campbell (1955). They examined accuracy and bias among seventh- and eighth grade African American and White students' perceptions of one another. All students predicted the score that each other student would receive on an upcoming test. How accurate were these students' perceptions of one another as individuals? They only reported results for accuracy in perceptions regarding the African American students, although they reported results separately by race of the perceiving students. Those results showed typically large accuracy effects. The correlations between predicted and actual scores were .56 for the White student perceivers and .47 for the African American student perceivers.

How accurately did the students predict race differences in test scores? Clarke and Campbell (1955) did not perform this exact analysis. However, their results did show that, on average, African American students predicted African American students' performance accurately (no discrepancy from perfection), whereas White students slightly underestimated the performance of the African American students (by about two tenths of a standard deviation). Unfortunately, Clarke and Campbell (1955) did not report analyses regarding accuracy in perceptions regarding White students. Therefore, all that we can conclude is that White students slightly underestimated the performance of the African American students, whereas the African American students did not.

Did relying on race increase or reduce the accuracy of perception of group differences? Again, because they did not report results for perceptions regarding White students, this cannot be determined from their study.

Cohen (1981). This study was described in detail in Chapters 5 and 9 but needs to be summarized briefly here because it is one of the few studies of stereotypes and person perception that provided data relevant to whether relying on stereotypes increased or reduced accuracy. Cohen (1981) examined people's memory for a videotaped conversation between a man and a woman where the woman was identified as either a waitress or librarian. How accurate were people's perceptions of the woman? Pretty accurate: On average across the two studies, they accurately remembered about 70% of the details of the conversation.

How accurate were people's perceptions of differences between the librarian and waitress? Well, in this study, *there were no real differences*, because they were the same person having the same conversation. However, across the two studies, perceivers consistently remembered the target as having 5% to 10% more stereotype-consistent attributes than stereotype-inconsistent attributes. This seems to imply that, had there been real differences, the perception of those differences would likely have been exaggerated (with librarians being remembered as more "librarianlike" than they really were and with waitresses being seen as more "waitresslike" than they really were). But wait . . .

Did relying on stereotypes increase or reduce accuracy? It *increased* accuracy. Having the label *before* viewing the tape (i.e., when the label had a chance to influence perception) increased the accuracy of people's memories by 7% compared to having the label only *after* viewing the tape (i.e., when the label had no chance to influence perception—see Chapter 9 for more details about this study). Stereotypes did influence, even bias, memory. But this is one of the earliest studies to show that relying on a stereotype *increased* the accuracy of people's judgments

Macrae et al. (1994). The authors of this series of three studies framed them almost entirely as testing the “cognitive miser” model discussed in Chapter 1. They suggested that stereotypes function to allow quick and “efficient” processing and simplification of a complex social world. Fortunately for this chapter they did so by testing how well people remembered the traits of particular individual targets. In other words, although Macrae et al. (1994) did not frame their study this way, it is one of the few published articles on stereotypes and person perception capable of assessing whether relying on a stereotype increased or reduced accuracy in person perception.

Accuracy (although they never used that stigmatized [see Chapter 10] term) was assessed as follows. Perceivers' task was to attempt to remember as many traits about individuals as possible (while simultaneously engaging in another task). Names of four individuals were always given (via computer screen). Half the time, targets were also labeled with a stereotype. Macrae et al. (1994) also provided a list of 10 traits that described each individual. So, some people's task was simply to attempt to remember the traits of Nigel, Julian, John, and Graham; other people's task was to remember the traits of Nigel—doctor, Julian—artist, John—skinhead, and Graham—estate agent.

So, did people do better with or without the stereotype? With the stereotype, hands down. In Study One, on average, they remembered over six traits correctly when they had a stereotype; they remembered less than 3.5 without the stereotype. In Study Two, even though the stereotype label was presented subliminally (i.e., so quickly that people were not even aware there was a label—for 3/1,000 of a second), people *still* remembered about five of the traits when they had the stereotype, but only three without the stereotype. In Study Three, they remembered about 10 traits correctly with a stereotype label, but only 7 without. For the statistically inclined, in all three studies, these differences were significant. In all three studies, this increase in accuracy occurred more for stereotype-consistent traits than for stereotype-neutral traits (e.g., the label increased the likelihood of people remembering a trait like “aggressive” for John—skinhead more than it increased their likelihood of remembering a trait like “modest” for John—skinhead).

Nonetheless, although this study was not at all framed as “replicating” Cohen's (1981) work, with respect to testing whether stereotypes increase or reduce accuracy of person perception, it did, in fact, do so. Actually, neither study was at all framed as testing the hypothesis that stereotypes increase the accuracy of person memory. To do so would likely have been the death knell of both studies—neither would likely have been published and almost certainly not in the *Journal of Personality and Social Psychology*, the most prestigious, highly cited, and widely read of all social psychology research journals. Nonetheless, *had* either study been framed as testing the hypothesis that stereotypes increase the accuracy of person memory judgments, the researchers would have been compelled to include that (in their studies) that hypothesis was confirmed.

Brodt and Ross (1998). The utility of an accurate stereotype was also demonstrated by Brodt and Ross (1998). College students made predictions about the behaviors and preferences of other college students who lived in one of two dormitories (one which had a campus reputation as a “preppie” dorm, the other as a “hippie” dorm). The students in the preppie dorm were widely seen as politically conservative, wealthy, and conventional. The students in the hippie dorm were widely seen as politically left wing with unconventional practices and preferences. Perceivers (other students who did not live in either dorm) viewed photographs of individual targets, were informed of each target’s dorm, and then made predictions about each target’s behaviors and attitudes (e.g., do they prefer eating at a vegetarian restaurant or a hamburger joint?). Perceivers’ predictions were then compared to the targets’ self-reports on these same preferences and attitudes.

So, how accurate were people’s predictions? Their accuracy depended on whether they relied on or ignored their stereotypes. When perceivers predicted targets to be consistent with their dorm (for a preppie dorm resident to have preppie attributes or for a hippie dorm resident to have hippie attributes), 66% of their predictions were correct (they matched the targets’ self reports). When perceivers jettisoned their dorm stereotypes and predicted targets to be inconsistent with their dorm, 43% of their predictions were correct. Relying on the preppie/hippie dorm stereotypes enhanced the accuracy of person perception predictions. Although they did not report results concerning the accuracy of perceived group differences, it is clear that relying on the stereotype increased the accuracy with which people perceived individuals.

Gosling, Ko, Mannarelli, and Morris (2002). In an entirely different context, this study demonstrated that, in general, perceptions of gender and racial differences are mostly accurate. One of Gosling et al’s (2002) main purposes was to determine how accurately people form impressions of individual targets’ extraversion, agreeableness, conscientiousness, emotional stability, and openness to experience², based on the information observable from either their office spaces (Study One) or bedrooms (Study Two).

Consistent with the bulk of the research reviewed earlier in this chapter, people judged individuals overwhelmingly on the basis of their personal characteristics (effects ranging from about .5 to about .8³). They did, however, also sometimes perceive gender and race (White vs. Asian) differences in personality (effect sizes for perceived differences averaged about .2 for gender and .3 for race). Were these perceived differences inaccurate? Well, it depends on how one performs analyses to answer that question.

Gosling et al (2002) examined this issue as follows: If and only if people perceived a statistically significant difference between genders or races, Gosling et al (2002) then examined whether there was also a significant difference in the same direction on criteria. Using this method, both Studies One and Two showed that people accurately perceived gender differences in emotional stability (men were perceived as, and were, higher than women). They misperceived gender differences in agreeableness. Women were seen as more agreeable, but there was a nearly significant difference in actual agreeableness in the other direction (i.e., men were more agreeable on the criterion) in Study One and no significant difference in Study Two. There were not enough non-Whites of any particular group to perform racial stereotype accuracy assessment in Study One. However, Study Two showed that people accurately perceived Whites as more open to experience than Asians, but inaccurately perceived Whites as more extraverted, more emotionally stable, and less agreeable than Asians.

So far, this looks like a pretty mixed bag with more inaccuracy than accuracy. Gosling, et al. (2002), interpreted these results as consistent with stereotypes influencing judgments (in some cases to be accurate, in most to be inaccurate), although they were also appropriately cautious in interpreting this. That is because their data do not conclusively bear on the issue of stereotype use. One may accurately perceive differences between two groups either because one relies on an accurate stereotype or because one relies entirely on relevant individuating information, and the groups differ in their average levels on that individuating information. Inaccurately perceiving differences between groups, however, strongly suggests reliance on an inaccurate stereotype, though it could also indicate misuse of individuating information that varies between groups.

In part, however, because Gosling et al. (2002) were trying to reach conclusions about stereotype reliance, the analyses they performed are, in fact, quite limited with respect to assessing the (related but different question of the) accuracy of the perceived differences. Next, therefore, I discuss why their analyses could not (and were not intended to) address the issue of the accuracy of the perceived differences between groups.

First, Gosling et al. (2002) simply ignored all the situations where people accurately perceived no differences between the groups. To assess reliance on stereotypes, this may have been a reasonable strategy; but to assess accuracy of people's beliefs about groups, accurately perceiving no difference between groups when there is no real difference should count as much as accurately perceiving a difference when there is a real difference. Second, presence or absence of "statistical significance" is a very crude measure of accuracy, sort of like estimating height with two categories, "tall/not tall" versus estimating height in feet and inches.

Fortunately, in one fell swoop (see Chapters 16 and 17), these limitations can be easily eliminated and a thorough and sensitive analysis of accuracy can be conducted: Simply correlate the perceived differences (or lack thereof) with the actual differences (or lack thereof). This allows for: 1) Full use of all the data (i.e., the analyzed data are not restricted to only when people perceived differences); and 2) Makes full use of all the quantitative nuances in the data (i.e., how much difference was perceived and occurred, rather than the cruder, qualitative is/is not statistically significant method used in the original article).

For gender in Study One, this correlation between perceived and real differences was nearly 0 ($r=.05$), suggesting no accuracy at all. However, this was almost entirely a function of the one agreeableness judgment that Gosling et al.'s (2002) perceivers clearly got wrong. After removing that one clearly inaccurate outlier, the correlation jumps to $r=.52$. So, for the other four personality characteristics, perceptions of sex differences were quite (though even then not perfectly) accurate.

Study Two did not have any extreme outlier, and, including all the data showed that perceived gender differences correlated $r=.93$ with the real differences. Perceived race differences correlated $r=.85$ with actual race differences. These are stunningly high levels of accuracy.

Overall, therefore, these results are broadly consistent with the Stereotype Rationality Hypothesis. People judged others primarily on the basis of their personal characteristics. Of course, information in one's living and working spaces is not completely diagnostic with respect to personality, so absent other more definitive information, it is, in fact, reasonable and rational for people to base their judgments, in part, on stereotypes. And, for the most part, people's perceptions of demographic differences (or lacks thereof) corresponded quite

well with the actual pattern of differences (or lacks thereof). Although whether this resulted from reliance on accurate stereotypes, or from reliance on diagnostic individuating information that varied between groups (or both) is not completely clear from their data. Either way, however, people usually ended up in the “right” (i.e., mostly accurate) place. Of course, the gender stereotypes regarding agreeableness found in the first study were highly inaccurate, indicating that not all of Gosling et al.’s results necessarily supported accuracy or the Stereotype Rationality Hypothesis.

Jussim et al., (1996); Madon et al. (1998). Jussim et al. (1996) and Madon et al. (1998) examined the accuracy of teacher expectations in sixth- and seventh grade classes, respectively. Both assessed teachers’ perceptions of their students’ performance, talent, and effort at math about 1 month into the school year. Accuracy was assessed in the following manner. First teachers’ perceptions of group differences were assessed by correlating teachers’ perceptions of individual students with the students’ race, sex, and social class. This correlation indicated the extent to which teachers systematically evaluated individuals from one group more favorably than individuals from another group. Next, actual group differences in performance, talent, and effort were assessed by correlating individual students’ final grades the prior year (before teachers knew the students), standardized test scores, and self-reported motivation and effort with students’ race, sex, and social class. The teachers’ accuracy was assessed by correlating the teachers’ perceived differences between groups with the groups’ actual differences.

How accurate were teachers’ perceptions of individuals? In both studies, accuracy was substantial: Consistent with research reviewed earlier in this chapter demonstrating the power of individuating information, the primary influences on teacher perceptions were students’ prior performance and motivation (multiple correlations of about .5 with teacher perceptions).

How accurate were teachers’ perceptions of group differences? In both studies, teachers’ perceptions of demographic differences regarding performance and talent were quite accurate. Perceived sex, social class, and race differences in performance and talent generally closely corresponded to the actual differences. For example, teachers perceived girls as performing slightly higher than boys (correlation between student sex and teacher perceptions of performance equaled $-.10$), and girls did, in fact, perform slightly higher than boys (correlation between sex and grades the prior year was $-.10$). In fact, out of the 18 teacher accuracy outcomes (sex, race, social class by performance, talent, and effort by two studies), only two—teacher perceptions of sex differences in effort in both studies—were substantially inaccurate (teachers perceiving girls as trying harder than boys, r s = $-.16$ and $-.24$, respectively, when, in fact, there were no differences in effort).

Madon et al. (1998) directly assessed the overall accuracy of teacher perceptions of differences by correlating the nine perceived differences with the nine actual differences. That correlation was $r = .71$. However, when the one clearly inaccurate outlier was removed (teacher perceptions of effort), the correlation between perceived and actual group differences increased to $r = .96$.

Did relying on stereotypes increase or reduce teacher accuracy? First, in both studies, there was no evidence that teachers even relied on ethnic or social class stereotypes. Both found that, when controlling for individuating information (motivation, achievement, etc.), student social class and race/ethnicity had little or no effect on teacher expectations.

Thus, teachers jettisoned their social class and ethnic stereotypes when judging differences between children from different social class and ethnic backgrounds. They did perceive differences between these demographic groups of children, but it was not because they relied on stereotypes—it was because they relied on the individuating information which itself reflected the very real differences between the demographic groups. Although this finding is in many ways laudable, teachers relying entirely on individuating information does not help address the question of whether relying on a stereotype increases or reduces accuracy.

Both studies, however, found that sex stereotypes biased teachers' perceptions of boys' and girls' performance and effort (standardized regression coefficients of .09 and .10 for performance, and .16 and .19 for effort, for Madon et al. and Jussim et al., respectively). In both studies, teachers perceived girls as performing higher and exerting more effort than boys. Because these effects occurred in the context of models controlling for individuating information, they are best interpreted as stereotypes influencing teacher perceptions—bias effects, in traditional social psychological parlance.

Did these sex stereotyping bias effects increase or reduce the accuracy of teachers' perceptions? They did both. The results regarding effort provided evidence of bias that reduced accuracy. There was no evidence that girls exerted more effort than boys. Therefore, the influence of student sex on teacher perceptions of effort—that is, teachers' reliance on a sex stereotype to arrive at judgments of effort—led teachers to perceive a difference where none existed. This is an empirical demonstration of something that, logically, has to be true and which was pointed out earlier in this chapter. Relying on an *inaccurate* stereotype when judging individuals can only harm one's accuracy.

In the case of performance, however, relying on the sex stereotype effect increased teacher accuracy. The real performance difference, as indicated by final grades the prior year, was $r = .08$ and $r = .10$ (for the 1996 and 1998 studies, respectively, girls received slightly higher grades). The regression model producing the “biasing” effect of stereotypes yielded a “bias” that was virtually identical to the real difference. In other words:

The small independent effect of student sex on teacher perceptions (of performance) accounted for most of the small correlation between sex and teacher perceptions (of performance). This means that teachers apparently stereotyped girls as performing slightly higher than boys, independent of the actual slight difference in performance. However, the extent to which teachers did so corresponded reasonably well with the small sex difference in performance. In other words, teachers' perceptions of differences between boys and girls were accurate because teachers relied on an accurate stereotype. (Jussim et al., 1996, p. 348)

The same conclusion, of course, also characterizes the results for the 1998 study.

Overall, as in the Gosling et al. (2002) study, these results from naturalistic research conducted in a setting of considerable importance in the real world of education were also broadly consistent with the Stereotype Rationality Hypothesis. Teachers relied overwhelmingly on the abundant individuating information. With respect to race and social class, they did not rely on stereotypes at all. For perceptions of performance, there was a very small sex stereotyping bias effect, but this effect increased, rather than reduced accuracy. Of course, the results showing that sex stereotypes biased and reduced accuracy in judgments of effort

are not consistent with the Stereotype Rationality Hypothesis and constitute an empirical warning against interpreting the research (or my claims) as suggesting that stereotypes are always perfectly rational or perfectly accurate.

Stereotypes and Person Perception: Conclusions

The prior chapter ended by summarizing the tactical retreat commonly taken by those wishing to acknowledge the existence, yet dismiss the importance of, the evidence on stereotype accuracy: “Yes, but what is really important about stereotypes is how they lead to biased judgments regarding individuals.” Fortunately (for real people, if not for the psychologists emphasizing the power of stereotypes), the evidence overwhelmingly shows that stereotypes do not lead to very large biases in person perception. Stereotypes biasing person perception is one of the smallest effects in all of social psychology; reliance on individuating information is one of the largest effects in all of social psychology; useful individuating information often eliminates stereotyping and nearly always reduces it by a great deal; and even ambiguous or useless information sometimes reduces stereotyping. Stereotypes can and do bias person perception judgments. But, like other expectancy effects, such effects tend to be small, fragile, and fleeting, rather than large, pervasive, and powerful.

Notes

1. That they are waiting for a train is itself a piece of individuating information, albeit a largely ambiguous one that is not very informative. This section describes why using both the individuating information and an accurate stereotype will usually enhance accuracy when people have ambiguous individuating information.
2. These five personality traits will be familiar to experts in social psychology and personality, and are known as the Big Five personality traits, because they have repeatedly emerged in empirical research on personality. Gosling et al (2002) used a common and well-validated questionnaire to assess both observer and self-reports on the Big Five.
3. For the statistically uninitiated, these are interpretable as correlations between gender or race and personality. For the statistically initiated, Gosling et al. (2002) actually reported η^2 s.

19 Stereotypes Have Been Stereotyped!

The Scientific and Social Value of Stereotype Accuracy Research

The last four chapters have shown that it is logically incoherent to define stereotypes as inaccurate, that it is unusual (but not unheard of) for stereotypes to be highly discrepant from reality, that the correlations of stereotypes with criteria are among the largest effects in all of social psychology, that people rarely rely much on stereotypes when judging individuals, and that, sometimes, even when people do rely on stereotypes, it increases rather than reduces their accuracy. Many scholars, scientists, and people of goodwill undoubtedly find these conclusions unbearable. They are justified nonetheless.

Stereotypes can be accurate. Some scholars and laypeople resist this conclusion, believing that crediting any accuracy to stereotypes is tantamount to endorsing bigotry. The opposite seems to me to be more likely true—that acknowledging the accuracy of some stereotypes provides the logical, definitional, and theoretical clarity needed to more effectively address prejudice and bigotry, and to more effectively investigate the nature, causes, and consequences of stereotypes.

Distinguishing accurate from inaccurate stereotypes. Not all stereotypes are accurate, and those that are inaccurate may be the most damaging. A special and important case is that of manufactured stereotypes, which are intentionally designed to despoil the reputation of particular social groups. A few notorious examples include 19th-century American stereotypes of indigenous peoples as uncivilized savages, stereotypes of civil rights workers as Communist fifth columnists, and the perpetual stereotype of Jews as seeking world domination. All these manufactured stereotypes served nefarious agendas, and all were (and are) patently false.

However, exposing the fallacious nature of these libelous stereotypes requires criteria and tools for assessing stereotype accuracy. These tools must be calibrated against a standard of authenticity, just as the tools for demonstrating counterfeit and fraud in art and business must. Whereas Jews do not seek world domination, it is not always absurd to believe that certain groups seek domination over, if not quite the whole world, at least large parts of it (consider, e.g., Rome, Nazis, Communists, Imperial Japan, and the Mongolian Khans). Without standards and methods for assessing (in)accuracy, it becomes impossible to reliably sort out valid beliefs from bogus beliefs.

Investigating the dynamics of stereotypes. Stereotypes are not static phenomena, but shift with circumstance, policy, social contact, and other forces. To what degree do stereotypes map these changes? How responsive are they to social shifts or to targeted interventions? Why do some stereotypes shift rapidly and others remain entrenched? Perhaps not surprisingly, if one makes the common assumption that stereotypes are inaccurate and answers these questions by assumption, one is not likely to even consider such questions, let alone provide answers to them. However, answers to some of these questions have indeed begun to be provided by researchers who make the alternative assumption, that stereotypes might be influenced by social reality (e.g., Diekmann, Eagly, & Kulesa, 2002; Oakes, Haslam, & Turner, 1994).

Generating a coherent understanding of both past and future research. The decades of research on the role of stereotypes in expectancy effects, self-fulfilling prophecies, person perception, subtyping, and memory are jeopardized if all stereotypes are regarded as wholly inaccurate. This past research will be haunted by a scientifically incoherent definitional tautology: That people who believe in stereotypes are in error because stereotypes are erroneous beliefs. On the other hand, accepting that stereotypes range in accuracy makes this past research coherent and allows for more edifying interpretations of past and future research, such as “people in X condition, or of Y disposition, are more likely to believe in, subscribe to, and maintain false stereotypes, whereas people in A condition, or of B disposition, are more likely to believe in, subscribe to, and maintain accurate stereotypes.”

In sum, accepting that stereotypes can sometimes be accurate provides the means to distinguish innocent errors from motivated bigotry, assess the efficacy of efforts to correct inaccurate stereotypes, and reach a more coherent scientific understanding of stereotypes. This proposition can advance the depth, scope, and validity of scientific research on stereotypes; help improve intergroup relations; and provide deeper and more well-justified insights into the nature of human psychology.

What Research on Stereotype Accuracy Does and Does Not Show

Because the term “stereotype” has so many pejorative connotations, and because it is so firmly associated with prejudice, discrimination, and injustice in so many people’s minds, I am going to (1) first clearly state many of the things this literature does not show, (2) state what it does show, and (3) describe many of the limitations to existing research on stereotype accuracy. I hope that doing so reduces the extent to which readers misinterpret my claims about what the research does show.

WHAT THIS RESEARCH DOES NOT SHOW

1. The research reviewed in Chapters 15 through 18 does not show that all stereotypes are always perfectly 100% accurate. No study has ever found this.
2. It does not show that prejudice and discrimination do not exist or are trivial and unimportant. Prejudice and discrimination are terribly important and can be terribly destructive. The research reviewed in this chapter has not addressed prejudice and discrimination; therefore, it has no direct implications for understanding prejudice and discrimination *per se*.

It does, however, raise the possibility that, despite suggestions that stereotypes are the “cognitive culprits” in prejudice and discrimination (e.g., Fiske & Neuberg, 1990; Fiske & Taylor, 1984, 1991; Myers, 2002; Stangor, 1995), stereotypes—which appear to reflect reality far more than they cause reality (compare the accuracy correlations in this chapter with, e.g., the expectancy effect sizes reviewed in Chapter 6)—may often play only a minor role in discrimination. Especially because stereotypes rarely correlate very highly with prejudice, it is becoming clear that factors other than stereotypes (such as hatred, domination, and conflict) may be the major sources of prejudice and discrimination (see Park & Judd, 2005, for a review).

3. It does not show that people correctly explain why group differences exist. Inasmuch as there is not widespread agreement among social scientists as to why group differences exist, it is not currently possible to assess the accuracy of most lay explanations for group differences. Social scientists (e.g., Hare-Mustin & Maracek, 1988; Jones, 1996) periodically object to the alleged naiveté of laypeople who, supposedly, assume group differences stem from biology (like the idea that stereotypes are inaccurate, the idea that people assume biological bases for group differences is often taken for granted without citation of evidence). In fact, however, the data suggest otherwise—the little research on people’s explanations for sex and race differences indicates that people believe that differences in socialization and opportunities are usually larger causes of race and sex differences in personality, interests, appearance, occupations, etc., than is biology (e.g., Martin & Parker, 1995; see Schneider, 2004, for a review).
4. It does not show how people arrive at their stereotypes. There is very little research on where stereotypes come from. Much speculative discussion emphasizes hearsay, family socialization, and the media (e.g., Allport, 1954/1979; Katz & Braly, 1933; Pickering, 2001). The extraordinary levels of accuracy shown in many of the studies reviewed in this chapter, however, do suggest another source is the primary basis of stereotypes—social reality.¹
5. It does not show that relying on a stereotype necessarily, or even usually, increases the accuracy of person perception judgments. Under certain circumstances, relying on a stereotype can indeed increase accuracy and has been shown empirically to do so. This, however, is a far cry from the claim that doing so always or inevitably increases accuracy.

WHAT THIS RESEARCH DOES SHOW

1. If stereotypes are beliefs about groups, defining them as inherently inaccurate has been falsified. If stereotypes are defined as inaccurate, they must always be inaccurate (see Chapter 15), and they must be almost entirely inaccurate. This is false.

2. An only slightly more modest claim or hypothesis, one that does not define stereotypes as inherently inaccurate, is that they are generally inaccurate. This also has been falsified. The scientific evidence provides more evidence of accuracy than of inaccuracy in social stereotypes. The most appropriate generalization based on the evidence is that people's beliefs about groups are frequently moderately to highly accurate, and only occasionally highly inaccurate. Consequently, social scientists should expunge all references to "inaccuracy" from their definitions of stereotypes (unless they expunge accurate beliefs about groups, and beliefs about groups of unknown validity, from both their definitions and from the research they consider to bear on stereotypes—which, as shown in Chapter 15, would logically require them to expunge nearly every scientific study of stereotypes that has ever been performed).
3. This pattern of empirical support for moderate to high stereotype accuracy is not unique to any particular target group or perceiver group. It has been found with racial/ethnic groups, gender, sororities, occupations, and college majors. There does, however, appear to be one exception to this pattern—political stereotypes seem to be less accurate than many other stereotypes.
4. This pattern of moderate to high stereotype accuracy is not unique to any particular research team or methodology. It has been found by a wide variety of American and Canadian researchers; by those using Judd and Park's (1993) componential methodology; by those using noncomponential methodologies; and regardless of whether the criteria are obtained through official government reports, meta-analyses, or the self-reports of members of the target group.
5. This pattern of moderate to high stereotype accuracy is not unique to the substance of the stereotype belief. It occurs for stereotypes regarding personality traits, demographic characteristics, achievement, attitudes, and behavior.
6. The strong form of the exaggeration hypothesis—either defining stereotypes as exaggerations or claiming that stereotypes usually lead to exaggeration—is dead. It has been killed by the data. Exaggeration does sometimes occur, but it does not appear to occur much more frequently than does accuracy or underestimation, and may even occur less frequently.
7. The exaggeration hypothesis—as a *hypothesis*—can still be retained. Exaggeration sometimes does occur. Perhaps some conditions systematically lead to exaggeration, whereas others systematically lead to underestimation. Understanding when stereotypes are more likely to exaggerate real differences, more likely to underestimate real differences, and more likely to be accurate is an important question for future research.
8. In contrast to their reputation as false cultural myths perpetrated by exploitative hierarchies to keep the oppressed in their places, consensual stereotypes were, by far, not only the most accurate aspect of stereotypes, not only massively more valid than nearly all social psychological hypotheses, but also stunningly accurate by any standard. Because bias and accuracy are not mutually exclusive, this does not necessarily "disconfirm" theories emphasizing bias in stereotypes. Nonetheless, correlations of .7 and higher are almost never systematically and repeatedly obtained in any area of social or psychological research—with studies assessing the accuracy of consensual stereotypes a rare and notable exception. Using Rosenthal's (1991) binomial effect size display to translate correlations into intuitively meaningful relationships

shows that correlations of .6 to .9 mean that consensual stereotypes are about 80% to 90% accurate. That level of accuracy is almost never reached in any other area of social life.

At both the individual and consensual levels of analysis, stereotypes—widely believed to be widely inaccurate—are more valid, and typically far more valid than most psychological hypotheses. Table 19–1 compares the frequency with which social psychological research produces effects exceeding correlations of .3 and .5 with the frequency with which the correlations reflecting the extent to which people’s stereotypes correspond to criteria exceed .3 and .5. Only 24% of social psychological effects exceed correlations of .3 and only 5% exceed .5. In contrast, *all* 23 of the aggregate/consensual stereotype accuracy correlations shown in Tables 17–1 through 17–3 exceed .3, and all but two exceed .5. Furthermore, 9 of 11 personal stereotype accuracy correlations exceed .3, and 4 of 11 exceed .5.

This is doubly stunning. First, it is yet another way to convey the impressive level of accuracy in laypeople’s stereotypes. Second, it is stunning that so many scholars in psychology and the social sciences are either unaware of this state of affairs, are dismissive of the evidence (for many of the unjustified reasons critically evaluated in Chapters 10 through 12 and 15), or choose to ignore it. When introductory texts teach about social psychology, they typically teach about the mere exposure effect (people like novel stimuli more after repeated exposure to it, $r = .26$), the weapons effect (they become more aggressive after exposure to a weapon, $r = .16$), that more credible speakers are more persuasive ($r = .10$), self-serving attributions (people take more responsibility for successes than failures, $r = .19$), parents encourage their children to adopt sex-stereotypic behaviors ($r = .21$), students with high self-esteem achieve more highly in school ($r = .21$), and romantic partners physically resemble one another ($r = .30$) (correlations all obtained from Richard, Bond, & Stokes-Zoota, 2003). These are all well and good and deserve the attention they receive in introductory psychology and introductory social psychology texts.

TABLE 19–1

Social Stereotypes Are More Valid Than Most Social Psychological Hypotheses			
	Proportion of Social Psychological Effects Obtained in Research ^a	Proportion of Consensual Stereotype Accuracy Correlations ^b	Proportion of Personal Stereotype Accuracy Correlations ^b
Exceeding .30	24%	100% (31/31)	86% (18/21)
Exceeding .50	5%	94% (29/31)	52% (11/21)

^a Data obtained from the Richard et al. (2003) review of meta-analyses including thousands of studies. Effects are in terms of the correlation coefficient, r .

^b From Tables 17–1, 17–2, and 17–3. Within parentheses, the numerator is the number of stereotype accuracy correlations meeting the criteria for that row (exceeding .30 or .50) and the denominator is the total number of stereotype accuracy correlations. Because Table 17–2 summarizes the results for five studies for McCauley, Thangavelu, and Rozin (1988), the .94 to .98 figure is counted five times. These numbers probably *underestimate* the degree of stereotype accuracy, because all single entries in Tables 17–1 through 17–3 only count once, even though they often constitute averages of several correlations found in the original articles, and because I did not use the r -to- z transformation (see endnote 1 in Chapter 17).

How much time and space is typically spent in such texts reviewing and documenting the much stronger evidence of the accuracy of people's stereotypes? Typically, none at all. Why not? Chapter 10 reviewed many of the reasons psychologists have been much more interested in error and bias than in accuracy, and Chapter 15 reviewed the historical emphasis on stereotype inaccuracy. Two additional reasons, however, are discussed here, in part because they may be particularly relevant to the emphasis on stereotype inaccuracy relative to accuracy. Specifically, I suspect that it is because of another delicious irony: Many social psychologists—who love to study errors and biases in laypeople's thinking—commit a logical fallacy.

THE PROCESSISTIC FALLACY

When social psychologists (see, e.g., any of those quoted in Chapters 5 and 15) actively emphasize and promote the idea that stereotypes are inaccurate, they almost always rely on scientific evidence to support this idea. How can this be possible, if the evidence so overwhelmingly demonstrates accuracy? The answer is that conclusions emphasizing error and bias rely on an entirely different body of evidence. Specifically, they rely on evidence regarding process, and this is evidence that may appear to, but does not actually, address accuracy.

To address accuracy, research must somehow assess how well people's stereotypes (or the perceptions of individuals) correspond with reality. The evidence that social psychologists typically review when emphasizing stereotype inaccuracy does not do this. Instead, that evidence typically demonstrates some sort of cognitive process, which is then presumed—without testing for accuracy—to lead to inaccuracy (the scientific alchemy by which this is accomplished was reviewed in Chapters 10 and 11). This is true for the pantheon of social psychological “errors and biases”—heuristics, illusory correlations, expectancy effects, group-serving attributions, system justification, etc.—all of which are real phenomena and are assumed (generally without test) to lead people to develop inaccurate beliefs about groups. And, in a limited sense, the claim that these phenomena and processes lead to inaccuracy is not necessarily false. They might cause inaccuracy, at least in the sense that the more that people engage in these processes, the less accurate their beliefs are.

Nonetheless, when experimental research uncovers some sort of flawed or biased process, the leap to the conclusion that, therefore, people's social perceptions are inaccurate is itself flawed and unwarranted. In fact, because this unjustified leap occurs so frequently, I think we need a new term to describe this flaw in scientific reasoning—the processistic fallacy. The processistic fallacy involves concluding that laypeople's beliefs must be inaccurate because researchers have discovered cognitive processes that the researchers believe to be flawed.

This is a fallacy for several reasons: (1) The process may not be as flawed as the researchers believe, and its degree of “flaw” cannot be assessed without assessing the validity or success of the judgments and decisions by people who do versus do not rely on this process (something social scientists rarely do); (2) even if the process is indeed flawed, in real life, people may rely on many other less flawed processes when making judgments and decisions; and (3) in real life, social reality often intrudes upon people's erroneous beliefs—that is, it provides feedback that permits people to recognize their initial beliefs were wrong and to alter them accordingly. So, again, we cannot know how flawed the outcome is—the judgment or decision—unless we evaluate its success, accuracy, validity, etc. (which is another thing social scientists emphasizing error and bias do not often do).

Thus, to review the evidence on the menagerie of errors and biases discovered by social psychologists and to conclude that, therefore, people's stereotypes are inaccurate is to commit the processistic fallacy. If we want to evaluate a baseball player's hitting ability, we cannot ignore his batting average, homeruns, etc., and instead determine the quality of his hitting ability exclusively by evaluating his swing. Similarly, evidence on accuracy cannot be inferred solely on the basis of research studying processes. Instead, to assess accuracy, research must compare people's perceptions of targets to criteria reflecting what those targets are actually like. The research that actually assesses accuracy, however, indicates that, despite whatever damaging effect allegedly flawed processes have on the accuracy of people's beliefs, both stereotypes and person perception judgments still often end up moderately to highly accurate.

WHY CONSENSUAL STEREOTYPES ARE USUALLY MORE ACCURATE THAN PERSONAL STEREOTYPES

Given the historical emphasis on the inherently inaccurate and evil nature of shared, cultural, consensual stereotypes, one of the most striking results emerging from the stereotype accuracy literature is the consistent extent to which consensual stereotypes corresponded stunningly well with real group differences. Psychological research almost never obtains results corresponding to correlations of .6 or higher, whereas the consensual stereotype accuracy correlations routinely exceeded .6. Why are these correlations so high?

The joy of averaging independent judgments. A book (Surowiecki, 2004) titled *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations* lays out a fairly straightforward way to understand this. Surowiecki starts his book by documenting case after case where the average judgment of a group is more accurate than the judgment of nearly all, and frequently all, of the individual members of that group. This is true when estimating the weight of an ox, the number of beans in a jar, or the location of a sunken submarine. Ever watch "Who Wants to Be a Millionaire"? This was a television game show in which the host asked contestants questions. They got to continue up the ladder (becoming eligible for more and more money at each rung of the ladder) as long as they answered the questions correctly. One incorrect answer and they lost.

One of the twists in the show was that people had several ways to request help if they were not sure of the answer for a question. One option was to poll the audience for their answer to the question. Ninety-one percent of the time, the majority picked the right answer, which is a helluva lot better than the contestants. This is also why it is very hard, when investing in stocks, to beat the market averages. If the average choice or estimate of a crowd is usually close to a dead-on bull's eye, then at any given time, most stocks will be fairly priced. One might make a decent return on a fairly priced stock, but one is unlikely to make a killing.

Why does this happen? When people make independent judgments, they bring a broad diversity of knowledge, background, insight, and intuition to bear on a problem. Some people's knowledge and background might lead them to be fairly close but to systematically overestimate the answer. Others' might lead them to be fairly close but to systematically underestimate the answer. With lots of people, the errors cancel out and all that is left is the part that is pretty close to the truth. Of course, some people's estimates may be almost completely random, idiosyncratic, and clearly wrong. However, random, idiosyncratic errors are,

by definition, just as likely to overestimate as to underestimate the real or best answer. An overestimate cancels out an underestimate, so that when you average them, the average will be closer to the truth than either estimate.² When people have at least some degree of knowledge or expertise relevant to a question, despite their individual imperfections, their group judgments are frequently going to be more accurate than their individual judgments.

Surowiecki (2004) also discussed how crowds or groups can go astray. When people's judgments are largely nonindependent, which can occur when social influence is high (e.g., during fads, when a strong or charismatic leader convinces many people to believe a certain thing, when a false cultural myth pervades a society, etc.), their errors and mistakes are no longer independent and random. They may systematically and consistently over- or underestimate some outcome. In such a case, the group's judgment is not likely to be very accurate. This type of process may occur in real life in situations where there is some sort of organized effort (e.g., by some sort of governmental or other institution) to distort the truth about some group (e.g., Nazi anti-Semitic propaganda, early American views of Indians and Africans as savages, etc.).

BUT, this only works if the average reflects the truth. So why are consensual stereotypes so valid? Well, the only way they can become so valid is if social reality has a systematic influence on individuals' beliefs about groups. This influence does not need to be large. If, however, social reality was completely unrelated to people's beliefs, those beliefs, even when aggregated, would not correspond with reality. Of course, the more highly the individual beliefs correspond with reality, the (even) more highly consensual beliefs will correspond with reality. That influence may be direct, obtained through personal experience with individuals from different groups, or indirect, obtained through family socialization, the mass media, education, etc. But one way or another, social reality appears to be the major influence on stereotypes.

So, the *Wisdom of Crowds* may help explain why consensual stereotype correlations are so stunningly high (frequently .8 and higher; see Chapter 17). The more important point, regardless of the explanation, is simply *that* consensual stereotype accuracy correlations are stunningly high. This result should constitute a dagger in the heart of (1) any modern definition of stereotypes as "inaccurate" and any implicit assumption of "inaccuracy" and (2) any perspective suggesting that social stereotypes are generally *false cultural myths*. The *shared* component of stereotypes, rather than being some sort of false cultural myth, is not only the *most accurate component of stereotypes* but also one of the very largest effects in all of psychology.

IMPORTANT LIMITATIONS

There are, of course, many important limitations to the existing work on the accuracy of stereotypes. First, the accuracy of two of the other major types of stereotypes—religion and social class—has, as far as I know, not received much attention. Although it is not clear that patterns of accuracy would differ for these types of groups, we will never know until the research is actually conducted.

Second, the existing research has overwhelmingly examined the stereotypes held by college students, largely because those samples are convenient. Is this important? Maybe. Suggesting it may not be that important has been the research by McCauley and colleagues

and by Clabaugh and Morling showing that the accuracy of noncollege groups is nearly identical to that of college students (see Tables 17–1 through 17–3).

Perhaps the “denial of difference” ideology (see Chapter 10) is more prominent on college campuses than elsewhere, and perhaps this ideology is part of what has created so much evidence of underestimation of real differences (although, again, McCauley’s research, often with nonstudent samples, also consistently finds evidence of underestimation). The only way to determine whether the accuracy of college students’ stereotypes differ from those of other people will be to conduct research assessing the accuracy of noncollege students’ stereotypes.

Third, there are many different types and aspects of accuracy, and few studies report results addressing all of them. Many studies have not made a strong distinction between personal and consensual stereotypes and report the results for only one or the other. Some studies focus primarily on discrepancies, others on correlations. Some focus on perceptions of variability across traits within a group; others focus on perceptions of differences between groups.

This is not a case where one is right and the other is wrong. They are all useful, and all provide information about different types or aspects of accuracy. Nearly all the studies I reviewed had the capability of addressing all of these types of accuracy but, for whatever reasons (not being aware of the different types of accuracy, not considering them important, journal space limitations, etc.), have not reported results for all of them. Ideally, more research in the future will provide more comprehensive assessments of the various types of stereotype accuracy.

Fourth, most research on stereotype accuracy has been conducted in the United States and Canada. Perhaps stereotypes in other countries are less (or more) accurate. The general hypothesis that social reality is the major source of many stereotypes probably holds true more strongly in highly educated societies and among more highly educated individuals, where people are more likely to bump into social reality and educators’ attempts to debunk false myths and stereotypes. In societies racked by poverty and ignorance, false beliefs about groups may be more likely to take hold.

Are Stereotypes Ever Highly Inaccurate?

THE EVIDENCE REVIEWED IN THIS BOOK

Evidence of major inaccuracy is rare but it is not entirely absent. First, even the studies that I have reviewed have provided some evidence regarding conditions under which stereotypes are less likely to be accurate. That can be summarized as follows:

1. Political stereotypes are not very accurate.
2. People are much better at judging differences between groups and at judging the rank order of attributes within a group than they are at judging the exact level of particular attributes within a group. In other words, the analyses assessing correspondence, which correlated people’s beliefs with group attributes or group differences, consistently found strong evidence of accuracy, whereas the analyses assessing discrepancies provided a much more mixed picture, including a fair amount of bull’s

eyes, a fair amount of near misses, and a fair amount of major inaccuracy. Even when they do not exaggerate or underestimate real differences, the evidence I reviewed showed that, often, people either consistently over- or underestimate the level of an attribute in a group. In short, although people's beliefs are often discrepant from reality to some degree, they are often quite good at capturing both the degree of differences between groups and the rank order of traits within groups.

3. On average, personal stereotypes corresponded well with groups' attributes (individual beliefs about groups correlated moderately to highly with criteria). Nonetheless, some personal stereotypes were highly inaccurate. Nearly all of the studies reporting personal stereotype accuracy correlations found at least some people with very low—near zero—correlations. Whether these are simply more or less random fluctuations and measurement error or whether some people are systematically more accurate than others is an important question for future research. Possible candidates would be intelligence (are smarter people more accurate?), education (are more highly educated people more accurate?), exposure to and experience with groups (the “contact hypothesis” [e.g., Allport, 1954/1979] has long suggested that contact with a group reduces prejudice, in part, by disconfirming erroneous stereotypes), nonverbal sensitivity (actually, Hall & Carter [1999] have already showed that people lower in nonverbal sensitivity hold less accurate sex stereotypes, but it would be useful to see if this pattern replicates), and ideology/motivated egalitarianism (several of the studies suggest indirectly that the more people are motivated to deny real differences, the less accurate their stereotypes, and Wolsko, Park, Judd, and Wittenbrink [2000] found this directly and experimentally).
4. The “egalitarian denial hypothesis.” I hereby introduce the “egalitarian denial hypothesis” (which I affectionately nickname as the *kumbaya hypothesis*) as a contrast to the exaggeration hypothesis that has dominated the literature, despite its lack of correspondence with much of the data. Kumbaya was a sort of “let's all love one another” song sung by hippies in the 1960s. The kumbaya hypothesis is that, in their attempt to be good, decent, unbigoted egalitarians, many people are motivated to deny real group differences. Because groups often do really differ on many attributes, such people, despite the benevolence of their intentions, perceive groups inaccurately. Specifically, such people see groups as differing less than they really do.

The kumbaya hypothesis differs from the exaggeration hypothesis in two major respects. The exaggeration hypothesis has generally been presented as a defining or common characteristic of stereotypes. In contrast, the kumbaya hypothesis is not intended to describe people in general. Instead, it predicts who will be more or less accurate: lower accuracy among people who deny group differences and higher accuracy among those who do not deny group differences. Who is likely to deny group differences? The kumbaya hypothesis predicts two groups: people on the far left of the political spectrum (consider, e.g., the Marxist emphasis on equality at the expense of freedom; e.g., Rokeach & Mezei, 1966) and people (regardless of ideology) highly motivated to be or appear egalitarian. These hypotheses should be tested in future research.

Second, the exaggeration hypothesis predicts that the main error in stereotypes is toward believing groups differ more than they do. In contrast, the kumbaya hypothesis identifies

people who are inaccurate because they believe groups differ *less* than they really do. Several different lines of research converge on support for the kumbaya hypothesis:

- The research by Ashton and Esses (1999) showing that people very low on right-wing authoritarianism—probably, very strong liberals—inaccurately underestimated real differences. And the finding that intelligence did not matter for this group strongly hints at the possibility that this denial of differences is motivated, rather than merely a reflection of ignorance or stupidity.
- The experimental research by Wolsko et al. (2000) showing that, when people are instructed to adopt a color-blind mindset, their stereotypes are less accurate than when instructed to adopt a multicultural mindset.

There is, however, one contrary study. The research by Hall and Carter (1999) showed that people who score high on universalism (i.e., the idea that we are all fundamentally the same) hold more accurate stereotypes than folks low on universalism. Given that the exaggeration hypothesis has endured intact for decades despite abundant evidence of underestimation, however, I think the kumbaya hypothesis is worth keeping around for a while and testing, despite this one study.

ARE NATIONAL PERSONALITY STEREOTYPES INACCURATE?

A large-scale study conducted in scores of countries all over the world found that there is also little evidence of accuracy in national stereotypes regarding personality (Terracciano et al., 2005). It is probably not surprising that people on different continents have little accurate knowledge about one another's personality (e.g., that Indonesians do not know much about, say, Canadians is not very surprising). However, somewhat more surprising is that people from cultures with a great deal of contact (various western European countries; Britain and the United States) also have highly inaccurate beliefs about one another's personality characteristics.

Although the Terracciano et al. (2005) study was impressive in scope and innovative in topic, it suffers from one of the limitations that excluded studies from the review conducted in Chapter 17. Specifically, the criteria samples were haphazard samples of convenience, rather than random samples obtained from target populations. The extent to which this explains their low level of accuracy is unknowable until research is conducted on the same topic that obtains criteria from random samples. Of course, it is also possible that national differences in personality are not readily detectable by laypeople, or perhaps such differences do not exist to any great extent. In general, why some stereotypes have such high levels of accuracy and others such low levels is currently unclear and is an important area of future research.

Furthermore, a recent reanalysis of the Terracciano et al. (2005) data (Heine, Buchtel, & Norenzayan, 2008) found evidence of accuracy consistent with results described in Chapter 18. Heine et al. (2008) argued that Terracciano et al. (2005) used an inappropriate criterion for assessing accuracy, although they suggested it was inappropriate for a different reason than I suggested. Specifically, the argument by Heine et al. (2008) was based on the well-known *reference group effect* (RGE)—the tendency for people to respond to self-report questions by using the standards of their own culture. They argued that the RGE undermines the validity

of self-reports for the type of cross-cultural comparisons used by Terracciano et al. (although the RGE does not undermine the validity of self-reports for within-culture accuracy assessments). As Heine et al. (2008, pp. 309–310) put it: “For example, one’s response to ‘I am not a very methodical person’ would hinge greatly on one’s understanding of the norms for being methodical. Because norms differ across cultures, the RGE systematically distorts cultural differences.” Furthermore, they argued, this is far more likely to be a problem for self-reports of personality than it is for “perceptions of national character”—the term used for “stereotypes” in the original Terracciano et al. (2005) paper—because people are more likely to realize they need to use different standards when evaluating people outside their own culture.

Regardless of argument, however, the issue (as usual) is data. To this end, Heine et al. identified five objective or behavioral measures that, they argued, should reflect “conscientiousness”—one of the personality characteristics assessed in the original study. Those measures were walking speed of 70 pedestrians unobtrusively measured in major cities, postal workers’ speed (operationally, how long it took a researcher to buy one stamp), the accuracy of the clocks in each of 15 randomly selected banks in the same city, gross domestic product, and longevity (well-established as a correlate of conscientiousness). They then correlated the average perceptions of national conscientiousness with each of these five criteria, in each of two ways, thereby producing 10 correlations. Nine of the 10 correlations ranged from .40 to .74; one was near zero. The average correlation of stereotype with criteria was .61.

These are, of course, consensual correlations. Unfortunately, Heine et al. (2008) were unable to obtain from Terracciano et al. (2005) the data on individual respondents’ stereotypes—which would have been necessary to assess personal stereotype accuracy. Nonetheless, these results are noteworthy in that they replicate almost perfectly the general patterns summarized in Chapter 17 (see Tables 17–1 through 17–3 and Table 19–1).

A BRIEF COMMENTARY: MORE BIAS IN FAVOR OF BIAS WITH RESPECT TO THE TERRACCIANO ET AL STUDY

In 2005, after the Terracciano study was accepted for publication but before it was actually published, I—along with scores, perhaps hundreds, of other psychologists—received an e-mail from a very famous and prestigious Ivy League psychologist alerting us to the publication of this “important” paper. Here I merely note several additional observations. I have never received from this psychologist—or, indeed any other psychologist—a mass notification of the importance of:

- the Heine et al. article demonstrating that there was far more evidence of accuracy in the Terracciano et al. respondents’ perceptions of national character than Terracciano concluded;
- any of the over 20 articles published from 1978 to 2005 demonstrating high accuracy in many social stereotypes (see Chapter 18 for the full review); or
- broad review articles concluding that the evidence shows stereotypes are often far more accurate than they are usually given credit for (e.g., Jussim et al., 2009; Ryan, 2002).

Apparently, at least in many social psychological circles, evidence of inaccuracy is so noteworthy that it warrants an e-mail alert. Evidence of accuracy is, in contrast, apparently viewed

as not worthy of the same sort of attention. As long as this sort of bias in favor of bias prevails, psychological scholarship will continue to create the distorted impression that people's social beliefs are far less valid and rational than they actually are.

What I am calling for here is not for my field to tout people as perfectly rational, but for a modicum of balance. Given that the bias side of the equation is well recognized, this constitutes a call for also recognizing the evidence of high accuracy. If, after 90 years of proclaiming the inaccuracy of stereotypes to the world, the evidence comes in showing that, overwhelmingly, people's beliefs about groups are moderately to highly accurate, can we really just say "never mind"? Or do we owe it to the public, to our students, to our field, and to ourselves to own up to the data and acknowledge that, however much our egalitarian goals motivate us to want to proclaim to the world the inaccuracy of stereotypes, in fact, dozens of studies now show that the beliefs about groups frequently held by laypeople are usually quite accurate?

SPECULATIONS ON OTHER CONDITIONS OF INACCURACY

Studies not reviewed here because they assessed people's beliefs about groups and then used as criteria the self-reports of haphazard samples of members of the target group (Allen, 1995; Martin, 1987) consistently find more evidence of what those researchers interpret as "inaccuracy." The disconnect between the stereotype and criteria, however, renders such results difficult to interpret.

The existence of so few clear and strong demonstrations of widespread stereotype inaccuracy does justify the conclusion that "Research on the accuracy of stereotypes usually finds evidence of moderate to high accuracy, and only rarely finds evidence of low accuracy." It does not, however, necessarily justify concluding that stereotypes are hardly ever inaccurate. Perhaps researchers have just not yet looked in many of the right places or right ways for stereotype inaccuracy.

Less accuracy in less affluent, less democratic societies? For example, education and mass communication levels are so high in the United States and Canada, where most of the stereotype accuracy research has been conducted, that perhaps, in general, people are more exposed to social reality there (and, probably, in other western democracies) than in many other places around the world. Perhaps the propaganda of demagogues in authoritarian regimes helps perpetuate and exacerbate inaccurate stereotypes. Stereotypes have indeed often been exploited and perpetuated in the service of ideologies justifying oppression (Jost & Banaji, 1994; Sidanius & Pratto, 1999), and it seems likely that stereotypes are more inaccurate in highly oppressive societies. The Jim Crow American South, South Africa under apartheid, the Indian caste system, and the Nazis' racial beliefs are a few examples that come readily to mind.

Unfortunately, because the powers-that-be under such systems are not likely to be open to challenges to their authority, it will probably be very difficult to perform studies of stereotype (in)accuracy in such contexts. If it is difficult to perform research in the contexts most likely to produce stereotype inaccuracy, the scientific literature will be skewed toward providing more evidence of stereotype accuracy than is actually true of people in general, around the world. I have no answer to this problem, except, perhaps, to urge the support of efforts to nonviolently bring the openness of liberal democracy, protection of individual rights, and

benefits of higher education to the four corners of the earth. But that is an issue well beyond the scope of this book.

Stereotypes of stereotypes. Some of the evidence reviewed in Chapter 17 also suggests that people's theories of oppression may lead them to inaccurate beliefs about groups. For example, the research by Beyer (1999) and Diekmann, Eagly, and Kulesa (2002) strongly suggests that a general belief that is, broadly speaking, true (women suffer discrimination) can lead to inaccurate sex stereotypes. Specifically, people erroneously underestimated the proportion of women in masculine majors and underestimated their GPAs (in Beyer's study), and also underestimated men's attitudinal support for political positions favoring women and women's rights (in the Diekmann et al. study).

This at least raises the possibility that other beliefs involving theories of oppression might be similarly inaccurate. Perhaps people underestimate Whites' support for Black civil rights and equality. Perhaps people overestimate the extent to which older people are viewed as frail or befuddled. Given the extent to which universities in particular attempt to sensitize students to histories of oppression and injustice, one can imagine a very broad array of hypotheses predicting stereotype inaccuracy based on the premise that people's stereotypes will presume that others support oppression and inequality more than they really do.

Consistent with this analysis, one of the clearest demonstrations of stereotype inaccuracy and exaggeration assessed *people's stereotypes of others' stereotypes* (Rettew, Billman, & Davis, 1993). For example, in one study, Rettew et al. (1993) had men estimate the percentage of men or women who fit several descriptions (e.g., prefer shopping over attending a sporting event, cry at a sad movie, etc.). They then had women estimate *what men would say* when asked to rate men and women on these same descriptions (this is women's stereotypes of men's sex stereotypes).

The results could not have been clearer. On average, there was a 29% sex difference in men's actual estimates. ***But the women estimated that men would expect a 45% difference.*** Women exaggerated men's gender stereotypes. Similar results occurred for business students' ratings of non-business students' stereotypes about business versus non-business students, and for a range of regional stereotypes (northeastern, southern, and Californian).

That is, people's beliefs about *other people's stereotypes* clearly seem to be exaggerated. Now juxtapose the (1) prevalence to this day of scholarly emphases on stereotype inaccuracy with (2) the overwhelming evidence of some, and sometimes a great deal of, accuracy, rationality, and reasonableness in stereotypes and stereotyping. Doing so, it is very hard not to reach the conclusion that many social science perspectives on stereotypes better characterize ***those same social science perspectives on stereotypes*** than they describe laypeople's stereotypes. Apparently, many social scientists' beliefs about stereotypes are better understood as reflecting the *psychological phenomenon* discovered in the research by Rettew et al. (1993) showing that people exaggerate others' stereotypes than as reflecting the *scientific phenomenon* of stereotype accuracy uncovered by the actual data. That is, many social science perspectives on stereotypes are ***exaggerated, inaccurate, and rigidly resistant to change in the face of relentless disconfirming evidence, and maintain their conclusions by virtue of a very selective focus on studies and findings that confirm the a priori belief in the irrationality and badness of stereotypes.*** It is hard not to see some of this as *irrational*; some clearly is logically incoherent. In other words, ***stereotypes have been stereotyped!***

Notes

1. Those of you reading this chapter by itself, without the rest of the book, may be thinking, “Oh, but as the abundant research on self-fulfilling prophecies shows, stereotypes may create that reality, rather than reflect it.” Unfortunately, however, this argument does not hold water. Even more unfortunately, you are going to have to read Chapters 4, 6, 10, 14, and 15 to fully understand why.

2. The statistical version of the answer is that aggregating judgments across judges increases the reliability of those judgments (Rosenthal, 1991). Increased reliability is the same as decreased random error. Random errors reduce correlations, so increasing reliability reduces random error, which increases correlations.

This page intentionally left blank

7 Conclusion

This page intentionally left blank

20 Important, Interesting, and Controversial Work on Accuracy, Bias, and Self-Fulfilling Prophecies That Did Not Fit Elsewhere

NO SINGLE BOOK can cover every study ever performed or discuss every possible issue or phenomenon related to some topic. This book has been no exception. Before trying to tie together all the material thus far presented, I first review some important phenomena or studies that belong in a book on relations between social beliefs and social reality, but which did not quite fit in any of the prior chapters.

Self-Fulfilling Prophecies

PROCESSES

My focus in this book has primarily been on the *outcome* of social interactions, rather than the processes by which, say, self-fulfilling prophecies occur. Such processes are, however, well-documented in the social science literature. When a self-fulfilling prophecy occurs, perceivers' expectations lead them to treat targets in accord with those expectations, and targets respond to that treatment in ways that confirm the originally erroneous expectation. In the case of expectancies involving warmth or friendliness, the self-fulfilling prophecy occurs because the target often reciprocates warmth with warmth and hostility with hostility (e.g., Snyder & Swann, 1978a; Snyder, Tanke, & Berscheid, 1977).

With teacher expectations, the process is similar but slightly more complicated. Teachers often act on their expectations by acting both more warmly to high-expectancy students and by holding them to a higher standard. They create a more pleasant learning atmosphere for

highs, provide more positive performance feedback to highs, provide more feedback (both positive and negative) to highs, give highs more challenging assignments, and provide highs with more opportunities to demonstrate mastery (e.g., Brophy, 1983; Harris & Rosenthal, 1985; Rosenthal, 1973). Highs respond to this benevolent treatment with increased learning, increased motivation, and, ultimately, achievement (Jussim, 1986).

MODERATORS?

Chapter 13 reviewed my own research on factors increasing or reducing the power of self-fulfilling prophecies in the classroom (as the statistically inclined know, these are called “moderators”). Others, too, have studied this general issue, in the classroom and in other contexts (see reviews by Jussim, Eccles, & Madon, 1996; Snyder, 1992). Self-fulfilling prophecies are more likely to occur when perceivers desire to arrive at a stable and predictable impression of a target (Snyder, 1992), when perceivers are more confident in the validity of their expectations (Jussim, 1986; Swann & Ely, 1984), and when they have an incentive for confirming their beliefs (Cooper & Hazelrigg, 1988). Self-fulfilling prophecies and perceptual biases are less likely when perceivers are motivated to develop an accurate impression of a target (Neuberg, 1989), when perceivers’ outcomes depend on the target (Neuberg, 1994), and when perceivers’ main goal is to get along in a friendly manner with targets (Snyder, 1992).

Self-fulfilling prophecies were strongest among the youngest students in the original Rosenthal and Jacobson (1968a,b) study, and they were stronger among first grade students than among students in third or fifth grade in a subsequent naturalistic study (Kulinski & Weinstein, 2001). These patterns suggest that younger children may be more malleable than older children and adults. However, a meta-analysis has shown that the strongest teacher expectation effects occurred in first, second, *and* seventh grade (Raudenbush, 1984). Further, the largest self-fulfilling prophecy effects yet reported were obtained in a study of adult Israeli military trainees (Eden & Shani, 1982). Although these findings do not deny the possibility that younger children are more susceptible to self-fulfilling prophecies, they do suggest that situational factors may also influence targets’ susceptibility to self-fulfilling prophecies.

One such factor may be new situations. People may be more susceptible to confirming others’ expectations when they enter new situations. Whenever people engage in major life transitions, such as entering a new school or starting a new job (including the military), they may be less clear and confident in their self-perceptions. Unclear self-perceptions render targets more susceptible to confirming perceivers’ expectations. This analysis may help explain the seemingly inconsistent findings regarding age. Students in first, second, and seventh grade, and new military inductees, are all in relatively unfamiliar situations. Therefore, all may be more susceptible to self-fulfilling prophecies.

New Self-Fulfilling Prophecy-Like Phenomena ---

STEREOTYPE THREAT

Stereotype threat occurs when a person’s fear of confirming a negative social stereotype of his or her group leads him or her to confirm that stereotype. It was first demonstrated among African American students, who performed worse on standardized tests when race was made

salient than when it was not, or when they believed it was an intelligence test (thereby implicitly raising cultural stereotypes about African Americans' supposed intellectual inferiority) rather than a test of cognitive skills (Steele, 1997). Similar patterns have been demonstrated for women and scores on standardized math tests (Steele, 1997). According to stereotype threat theory, this occurs because the anxiety associated with potentially confirming a negative stereotype of one's group distracts and/or debilitates the target, who then has fewer mental resources with which to focus on the task at hand (standardized test), resulting in a lower score.

At first glance, stereotype threat appears to be very much a self-fulfilling prophecy. Erroneous race (or gender) stereotypes lead African Americans (or women) to underperform on standardized tests, thereby "confirming" the stereotype, which, except for stereotype threat, would be false. And, in fact, research on stereotype threat has indeed often been discussed in this manner (see, e.g., Sackett, Hardison, & Cullen, 2004 for a review of such interpretations).

There is, however, reason to be cautious in interpreting such studies as self-fulfilling prophecy for two very different reasons. First, in contrast to the type of self-fulfilling prophecies discussed throughout this book, there is no "perceiver." In stereotype threat research, stereotypes constitute a "threat in the air," to quote Steele's (1997) famous phrase. Stereotype threat works because *targets* fear confirming stereotypes, not because any particular perceiver is imposing an erroneous belief on the target. The phenomenon, at least as it has been studied, exists entirely in the mind of the target. If there is no perceiver with an inaccurate expectation, although stereotype threat is still an interesting and important phenomenon, it is not really a self-fulfilling prophecy, at least not in the classic sense in which the term has been used for the last 50 years. Self-fulfilling prophecies refer to *someone's* false belief that leads to its own fulfillment.

A second problem with stereotype threat research is not the research itself, but the manner in which it has been interpreted—specifically, as allegedly showing that "if not for stereotype threat, African American standardized test achievement would be the same as that of Whites" (see Sackett et al., 2004, for a review). Unfortunately, despite how widespread this interpretation is, it is not valid. Although Steele and Aronson (1995, the paper first reporting the stereotype threat phenomenon) never actually made this claim, they do bear some of the responsibility for others misinterpreting their work in this manner. In that paper, the authors presented a graph that made it appear as if the typical African American/White standardized test score existed only under stereotype threat conditions, but disappeared once stereotype threat was removed (Steele & Aronson, 1995).

This appearance, however, did not match the reality. African American students underperformed compared to White students in all conditions. What Steele and Aronson (1995) found was that (1) in the control conditions, the typical race differences emerged, and (2) ***under stereotype threat, race differences in performance increased.*** This occurred because, under stereotype threat, African American performance declined whereas White performance was unaffected.¹ The typical race difference occurred under control conditions, and an even larger difference emerged under stereotype threat.

Therefore, as a phenomenon and process, stereotype threat is "valid"—stereotype threat undermines the achievement of African American students. As an explanation for race differences in standardized achievement tests, however, it fails because at no point in the study

did the African American students receive scores as high as those of the White students. Thus, even on its own terms (i.e., considering the threat “in the air” as some sort of cultural expectancy), stereotype threat largely fails as a self-fulfilling prophecy explanation for Black–White differences in academic achievement.

STEREOTYPE PRIMING AND SELF-FULFILLING PROPHECIES

Social psychology, its scientific aspirations notwithstanding, is, in part, a field of fads and “cascades” (the *Wisdom of Crowds* [see Chapter 19] term for how waves of social influence can lead large numbers of people to do essentially the same thing at the same time, or even believe the same erroneous thing at the same time [e.g., believing in 1999 that the stock market would provide 20% returns indefinitely]). In the social psychology of the 1940s and 1950s, the most favored topic was “attitudes”; in the 1960s, “cognitive dissonance”; in the 1970s, “attributions”; in the 1980s, it was “expectancies” and “schemas”; and since the 1990s, we have had a wave of research on “automaticity” and “priming” and all things “implicit” (e.g., Bargh & Chartrand, 1999; Devine, 1989; Greenwald & Banaji, 1995). Automaticity refers, in essence, to people doing things without much or even any conscious thought; “priming” refers to a variety of ways to “activate” people’s beliefs, motivations, stereotypes, prejudices, etc., usually outside of their own awareness; and “implicit” refers to the presence of beliefs, evaluations, prejudices, etc., that are outside of people’s conscious awareness.

What ties automaticity, priming, and all things implicit together is that they all occur largely or completely outside of conscious awareness. It is, by now, very clear that, although we are very good at coming up with plausible-sounding reasons for why we do many of the things we do, we are, in fact, at least sometimes and perhaps often, unaware of the actual reasons why we do what we do, believe what we believe, or evaluate others the way we do (e.g., Bargh & Chartrand, 1999; Nisbett & Wilson, 1977). Whether recent claims as to the power and prevalence of automaticity, priming, and things implicit (e.g., Bargh & Chartrand, 1999; Greenwald et al., 2002) reflect an actual state of affairs—that is, that people’s daily lives really are dominated by beliefs, attitudes, and motivations outside of their own awareness—or, like dissonance, attributions, expectancies, and schemas, merely researchers’ infatuation with the topic and phenomenon is probably something that won’t be known for another decade or more.

Nonetheless, out of this perspective has emerged a single, very interesting study of stereotype-priming and self-fulfilling prophecies. Chen and Bargh (1997) performed a semi kind of replication of the Word et al. (1974; see Chapter 4) classic study of racial stereotypes as a self-fulfilling prophecy. First, Chen and Bargh (1997) primed perceivers’ racial stereotypes by subliminally exposing participants to either African American or White faces. Then they had their perceivers play a game with targets, who were in a different room (a cover story made this seem appropriate for research purposes).

However, because they were studying priming/automaticity, they did not want people’s conscious beliefs or attitudes to interfere with their research. Therefore, they purposely did not have their White perceivers interact with an African American target/interviewee (as did Word et al., 1974). So, all targets/interviewees were White.

Results provided clear evidence of something they interpreted in self-fulfilling prophecy terms. Perceivers/interviewers who were subliminally primed with African American faces

(as opposed to White faces) acted in a more hostile manner to the targets and evoked greater hostility in return. The self-fulfilling prophecy interpretation is that “hostility” is a core component of stereotypes regarding African Americans, the priming mechanism activated this stereotype but only for those primed with African American faces, and this activated stereotype then influenced the course of the interaction, ultimately evoking greater hostility from the (White) targets.

Their conclusions emphasized the potential power of nonconsciously activated stereotypes to create self-fulfilling prophecies that undermine the performance of African Americans. Chen and Bargh (1997) also interpreted this to mean that, even in the absence of a manifestly erroneous expectation, the stereotype can create a self-fulfilling prophecy. That is, in contrast to much of the early work on self-fulfilling prophecies (see Chapter 6), Chen and Bargh did not induce an erroneous expectation by lying to their perceivers about the targets or by deceiving them in any way. All that Chen and Bargh (1997) did was to activate the stereotype.

But what does this really mean? In a situation like this, it is usually good, in my opinion, to start by sticking very close to the methods and results of the study. Perhaps most telling, with respect to understanding the potential relevance of this research of real interviews involving Whites and African Americans, was Chen and Bargh’s methodological decision not to have any African American interviewees. Why did they make this decision? Because conscious, controlled processes might have overridden their priming effects. Here is what they wrote about this issue (p. 549):

However, doing so [having perceivers interact with actual African American targets] would have precluded a test of our hypothesis. Because the perceiver would have been consciously aware of the African American race of the target, *conscious perceptual and behavior strategies on the part of the perceiver could not be ruled out as alternative explanations for any effects we would obtain*. . . . [T]he actual physical presence of an African American . . . would likely overwhelm the subliminal priming manipulation . . . [and] the priming effect would likely be mitigated entirely. (emphasis mine)

To this I reply—exactly! With respect to understanding discrimination, even if many Whites harbor nonconscious prejudices, *if they consciously and successfully fight those prejudices and behave in an approximately egalitarian manner*, there would be no self-fulfilling prophecy. I do not know whether this is what Chen and Bargh meant by this quote, but, regardless, this is what I think it *should* mean. Furthermore, if conscious processes are so much to fear (with respect to their experimental methods), perhaps human social behavior is not quite so dominated by automaticity as the current fads seem to suggest.

That Whites will at least sometimes actively fight against their own tendencies to be prejudiced or discriminate has been amply demonstrated (Devine, 1989; Norton, Sommers, Vandello, & Darley, 2006; Sommers & Ellsworth, 2000). For example, in a recent study of judgments regarding college admissions (Norton et al., 2006), participants reviewed college applicants with similar, very strong qualifications. There were two main applications: one with a 3.6 GPA and many difficult Advanced Placement (AP) courses and one with a 4.0 GPA and fewer AP courses. Half the time, race was left out, and the two applicants were judged about equally acceptable. When race was included, the Black applicant was chosen

over the White applicant for admission 78% of the time—a preferential selection effect size in favor of the Black applicant of about .60 (ranking it up there with stereotype accuracy as one of the largest effects in social psychology!).

At minimum, the possibility that many people would fight against their own biases leads me to wonder why Chen and Bargh (1997) did not run their study (or a replication) with both African American and White targets. Now, I cannot really fault them on this account. There is only so much any researcher can accomplish in any single study, and nearly all single studies have important limitations. Furthermore, much like my critique of the work on expectancy bias and self-fulfilling prophecy (Chapters 6 through 9), this critique does not challenge the internal validity of the Chen and Bargh study. It was a well-conducted study, nicely designed to test the hypothesis that priming White interviewers' racial stereotypes will undermine the performance of White interviewees (though, as usual, I think it would be wise to remember that this is a single study with, as far as I know, no published attempts at replication, and, especially, that many of the “dramatic” and highly cited expectancy-confirmation studies of the 1970s and 1980s could not be replicated).

Given the extraordinary limitations to this study, however, one might think that the conclusions based on it would be appropriately measured and cautious. In this context, let's also consider what Chen and Bargh (1997, p. 557) made of their results: “. . . once stereotypes are so entrenched in an individual as to become automatically activated, there is little probability that their biasing effects will be prevented. . .”

“Entrenched?” This is, at best, highly overstated. Perform a study that provides people with ZERO chance to recognize that they might act in a prejudicial manner (by not having people interact with African American targets), and then conclude that stereotypes are so entrenched that people cannot prevent bias? And purposely perform a study designed to prevent conscious control, and then conclude that biases probably cannot be prevented? Not to mention—how do these “entrenched stereotype biases unlikely to be prevented” square with the fact that the biasing effects of stereotypes on person perception are one of the *smallest* effects in all of social psychology (see Chapter 18)? Of course, it is possible that they are right to at least some degree, but until this type of research is replicated in such a manner that perceivers do have an opportunity to consciously override whatever unconscious biases they have, we will never know. And to draw such an extreme conclusion on the basis of a study with such severe limitations smacks of exactly the same type of infatuation with bias that has generally characterized social psychological perspectives on expectancies, self-fulfilling prophecies, and stereotypes more generally.

REJECTION SENSITIVITY

People in romantic relationships vary in how concerned they are about their partner's likelihood of rejecting them. Downey (Downey & Feldman, 1996; Downey, Freitas, Michaelis, & Khouri, 1998) dubbed this phenomenon “rejection sensitivity.” Rejection sensitivity seems to trigger a bona fide self-fulfilling prophecy that works as follows.

In most close relationships, people are not mushy warm lovey-dovey with each other all the time. Sometimes, perhaps, but not all the time. At any given time, the relationship may not be front and center to one person. He or she may be tired, distracted, tense about work, etc. Furthermore, people in even the best relationships do not always agree with one another and have some conflict.

Many people, especially in healthier relationships, understand this and treat it as “water off a duck’s back”—they do not feel their relationship is threatened by such events. People high in rejection sensitivity, however, are different. First, they are hypervigilant about detecting signs that their partner might be rejecting them (Downey & Feldman, 1996). Thus, an event that other people might treat as no big deal—for example, a missed return call, late arrival, one-word response to a question—people high in rejection sensitivity “detect” as evidence that their partner is rejecting them.

Thus, this heightened concern leads to an interpretive bias—all things being equal, people high in rejection sensitivity are more likely to interpret an ambiguous event as evidence of their partner’s rejection (Downey & Feldman, 1996). That alone, however, is not a self-fulfilling prophecy, which requires that this heightened concern about rejection lead to actual rejection. That is, however, what sometimes happens. The high-rejection-sensitivity partner perceives more instances of rejection; this leads to tension and conflict in the relationship (“why didn’t you return my call?!”), and the tension and conflict increase the likelihood of the relationship breaking up (Downey et al., 1998).

This research on rejection sensitivity provides some of the clearest evidence of a new type of real-world self-fulfilling prophecy found in the last 30 years. Downey’s studies have focused primarily on real couples involved in long-term romantic relationships, and she has used both laboratory and naturalistic methods. Although the self-fulfilling effects of rejection sensitivity on outcomes such as partner dissatisfaction and anger were typically modest (about 0.2), this program of research does not suffer from the types of problems and limitations that have typically plagued much other self-fulfilling prophecy research (such as intentionally providing perceivers with false information, failing to study naturalistic interactions, etc.). Nor has this program of research suffered from being overpromoted, oversold, and overstated in the same manner as has much other research on expectancies. As such, it is one of the most interesting newer avenues of self-fulfilling prophecy research to open up in some time.

Bias

Social psychology remains, in large part, a field of “bias finding” (see, e.g., Table 1–1 in the first chapter). This state of affairs has been true since at least the late 1970s and, with psychologist Daniel Kahneman’s 2002 receipt of the Nobel Prize in economics for work on biases in judgment and decision making, is not likely to die down any time soon. So, researchers are discovering new “biases” all the time.

This book is not about bias writ large, and there are way too many biases unrelated to expectancies to be reviewed in this concluding chapter. One recent set of studies, however, I find to be particularly intriguing, because they strongly imply that a racial bias emerges from a largely accurate racial stereotype (see also Chapters 10 and 18).

RACE “BIAS” IN JURY SELECTION

In one study, college students, law students, and attorneys (who were over 70% White and under 5% Black) were asked to take the role of prosecutor in a trial of a Black defendant (Sommers & Norton, 2007). The key question was whether they would exercise more

peremptory challenges to exclude potential Black jurors than to exclude potential White jurors. “Peremptory challenge” refers to attorneys’ right to exclude a limited number of potential jurors without having to explain or justify the reason for the exclusion.

This was an experiment, and their procedures ensured that, overall, the personal backgrounds of the potential Black and White jurors were identical. Nonetheless, the Black juror was rejected by 63% of the participants, the White juror by only 37%.

Is this race “bias”? Of course. Is it inaccurate? Sommers and Norton (2007) also assessed their participants’ beliefs about how likely the prospective jurors were to vote guilty. Sure enough, participants believed that the prospective Black juror was less likely to vote guilty. Remember, now, their participants were asked to assume the role of the *prosecutor*—that is, their *job* required them to make as strong a case as possible to as sympathetic a jury as possible, for the defendant’s *guilt*. So, if they believed one juror was less likely to vote guilty than another, rejecting that juror seems quite rational.

But was it accurate? Well, this was an experimental study, and there were no “real” jurors who actually voted on guilt. Thus, accuracy cannot be directly determined (of course, that also prevents anyone from concluding that Sommers & Norton’s [2007] participants were inaccurate). But it can be approximated by considering research on juror race and verdict. Interestingly, in prior experimental studies of mock jurors (people asked to read trial summaries and reach verdicts), the same Sommers (Sommers & Ellsworth, 2000) found that Black mock jurors were considerably less likely to conclude a Black defendant was guilty than were White mock jurors.

Thus, Sommers and Norton’s (2007) mock prosecutors did indeed seem to *correctly* assume that Black jurors would be less likely to vote to convict than White jurors. Furthermore, it was this assumption, rather than juror race per se, that led them to disproportionately exclude the potential Black juror. They performed a final analysis assessing whether these judgments of likelihood to convict mediated the effect of the potential juror’s race on the mock prosecutors’ use of peremptory challenge. Such judgments entirely mediated the effect of potential jurors’ race. In other words, there was *no significant effect* of potential juror race, after controlling for what may be a generally accurate belief that a Black juror would be less likely to convict a Black defendant than would a White juror.

So, as best as can be determined, Sommers and Norton’s (2007) mock prosecutors were, in fact, acting on a generally accurate stereotype regarding the role of juror race in verdicts. Black jurors may indeed be generally less likely to convict a Black defendant than are White jurors. Is this biased use of the peremptory challenge illegal? I am neither an attorney nor an expert at law psychology, so I do not really know. Both Sommers and Norton (who graciously commented on an earlier version of this section) assure me that it is illegal because it is illegal to rely on race. I am not so sure. It seems at least hypothetically possible that a good attorney could convince a judge that prosecutors acting as did Sommers and Norton’s (2007) participants were not relying on race. Given that there was no effect of race after controlling for perceived likelihood of convicting, it seems to me that a very good case could be made that attorneys acting in this manner were relying on individuating information—likelihood of convicting—and this relevant basis for peremptory challenges varied by race.

Regardless, even if my legally nonprofessional analysis is wrong and no judge would allow such a practice, the question remains as to whether such a practice *should* be allowed. If it is false to declare any use of race whatsoever to be unjustified bigotry because race is actually

relevant to judging jurors' likelihood of convicting, on what moral grounds should the use of race be prohibited? Why should some information that implies a bias for or against the prosecution be legal to use, whereas race, which, under some circumstances might also imply favoritism, should not be legal?

There may be answers to these questions, but because so much of the social sciences (and, indeed, the general public) seems to be in denial about the possibility that some stereotypic judgments may be based in reality, these types of questions do not usually even get asked.

Please note that nowhere in this chapter have I advocated legalizing use of race in peremptory challenges (or anywhere else). Neither in this chapter nor anywhere else in this book have I advocated any particular set of laws or policies. This book is about what logic and evidence have to say about the extent to which social beliefs create social reality and the extent to which social reality creates social beliefs, not about policy recommendations. Regardless, the research by Sommers and Norton points out the dangers of equating bias with inaccuracy and also highlights the potential conflict of two laudatory social goals in the courtroom: accuracy and nondiscrimination.

Without advocating any particular position, however, I do think the following statement about policy is justified. Whatever positions we ultimately decide to take on issues such as these, policies adopted on the basis of the kumbaya hypothesis, ignorance, or wishful thinking ("all stereotypes are inaccurate"; "judging individuals on the basis of stereotypes necessarily leads to less accurate judgments"; "there are no real race, sex, religion, or social class differences") are far more likely to be dysfunctional and produce unintended negative side effects than are policies adopted out of a fuller recognition of the extent to which accuracy and nondiscrimination may sometimes conflict (see also Dawes, 1994, for a similar analysis applied to affirmative action).

BIG BAD BIAS

People are subject to all sorts of biases that are interesting and important, but which are not the subject of this book. I have focused on biases that mislead people to interpret the social world as more consistent with their prior beliefs, stereotypes, and expectations than it really is. This requires first the obtaining of credible evidence on how the social world really is. In many situations, it is impossible to know with either certainty or even high probability how the world really is. Is abortion murder? Will allowing gays and lesbians to marry erode the moral fabric of society? Is there any one true religion? It is almost impossible to obtain much, if any, scientific evidence that bears on questions such as these. Thus, biases regarding these types of issues—for which there are no accuracy criteria—have been beyond the scope of this book. It is possible, therefore, that other biases—produced by religion, ideology, or other types of deeply held beliefs about things that are mostly outside of the realm of scientific study²—produce more extensive biases than do those addressed in this book.

However, before you, gentle reader, leave this book with the impression that "Aha! The places where the big bad biases can be found have been selectively ignored!" please keep the following in mind. First, although the biases reviewed in this book have not been exhaustive, in the realm of expectancy effects, when I have been selective, it has been to selectively focus on some of the most well-known, highly cited studies of expectancy-based or stereotype-based bias (e.g., Darley & Gross, 1983; Hastorf & Cantril, 1954; LaPiere, 1936; Rosenhan,

1973; Snyder & Swan, 1978a,b) and to show that *even those studies* often provide more evidence of accuracy and less of bias than is usually claimed, or that they have been subjected to repeated failures to replicate. For decades, and still often enough today, some of the most influential scholars and scientists in my field have proclaimed the biases that I *did* include in this book as powerful and pervasive. All that I suggest, therefore, is that the next time you come across some testament to the power of human error and bias, ask yourself a few simple, but pointed questions:

1. How big is that bias? What was the effect size?
2. Was the study purposely designed to make it difficult for people to be reasonable, rational, or accurate? Did it give them a reasonable chance to be reasonable? Did the study explicitly consider and acknowledge the ways in which the judgments that people did make might be reasonable, rational, and accurate under naturally occurring conditions?
3. Did the researchers explicitly present an analysis of what people *should* do to be as rational and accurate as possible? Or did they ignore this issue, leaving themselves the freedom to interpret *anything* people do as bias?
4. Has the study been replicated by anyone other than the original researchers, or their former students and protégés?
5. If the researchers did identify what a *perfectly rational or unbiased* judgment would be, did they then interpret any discrepancy from perfection as a testament to how biased people are *without comparing the degree of imperfection to degree of rationality or judgmental appropriateness*? Did that study of the role of stereotypes in person perception compare stereotype bias to reliance on individuating information, *or did it focus entirely on evidence of bias complete with a narrative discussion emphasizing how the results testify to people's inherently biased judgmental process* and play down or ignore reliance on individuating information? Did that study of self-fulfilling prophecies also compare the extent of self-fulfilling prophecies to accuracy, *or did it focus exclusively on self-fulfilling prophecy* and not even consider, let alone empirically assess, the possibility that the expectations might be accurate to at least some degree, and perhaps, to a greater degree, than they are self-fulfilling?
6. Did the study examine people making the types of judgments they usually do in real life, under conditions identical or at least close to those in real life?
7. Did the researchers forthrightly discuss limitations to their study, especially the potential to misinterpret its results as demonstrating a more powerful bias than it really found?

If the bias really is big, if the study gave people ample opportunity to be reasonable, if the researchers articulated a clear description of what it means to be rational and accurate and then showed people greatly deviated from this ideal, and if, furthermore, the bias was actually found under conditions pretty close to those found in real life—congratulations. Someone has actually found a big bias. Nothing in this book has suggested that doing so is impossible.

Nonetheless, and despite the existence of numerous testaments to their power, the biases that I did review in this book are, in fact, usually quite small. Perhaps someday someone will

provide evidence of a context or way in which expectancy-confirming judgmental and perceptual biases are genuinely powerful and pervasive, but until that time, the conclusions reached in this book—that such effects tend to be quite modest—will stand.

Accuracy

Despite the overwhelming emphasis on bias, pockets of research on accuracy have been conducted semiregularly, especially since the late 1980s. This research has examined accuracy in personality judgments, in beliefs about nonverbal behavior, in reading friends' and strangers' thoughts and feelings, and in many other contexts (e.g., Funder, 1987; Ickes, 1997; Kenny, 1994). In general, and with exceptions, this research shows that people are often at least moderately accurate in many different contexts. Although this book has focused on the accuracy of expectations in particular, expectations have to come from somewhere. The research reviewed here in chapters on teacher expectations (Chapters 3, 13, and 14) and on stereotypes (Chapters 15 through 19) has addressed some of the sources of expectations. Nearly all research on accuracy, however, can be viewed as providing information about the source of accuracy of expectations. Although a thorough review is not possible, some of the most interesting and creative research on accuracy is briefly reviewed next.

EMPATHIC ACCURACY

Empathic accuracy refers to the degree to which one person successfully infers the private, subjective thoughts and feelings of another person (Ickes, Stinson, Bissonnette, & Garcia, 1990). I find research on empathic accuracy striking for several reasons. First, to anyone steeped in the error and bias zeitgeist of modern psychology, especially at the time when accuracy was still largely "forbidden" (see Chapter 10), even attempting to assess this must have seemed like Don Quixote tilting at windmills. Accurately read another person's thoughts and feelings? Given that thoughts and feelings are not directly observable, to anyone who has bought lock, stock, and barrel the common social psychological story about the inherent ambiguity of so much of social reality, about "cognitive misers" being unwilling or unable to perceive social reality in all its richness and complexity, about the alleged extraordinary power of people's expectations, stereotypes, and preexisting schemas to distort reality, given the alleged power of phenomena such as the fundamental attribution error, this must seem like a fool's errand.

But it was no fool's errand. In an initial, exploratory study, Ickes et al. (1990) videotaped unacquainted strangers interacting for 6 minutes. The participants then reviewed the video, indicating their thoughts and feelings at specific points in the interaction. They then reviewed the video a second time to indicate not their own, but their partner's thoughts and feelings. Comparing these two (Partner A's perceptions of Partner B's thoughts and feelings to Partner B's self-reported thoughts and feelings) provided the measure of accuracy.

Results provided evidence of an extraordinary pattern of accuracy and inaccuracy. First, they created two separate domains of accuracy: content (what their partner was thinking) and valence (what their partner was feeling). Overall, participants were right about 22% of the time for content and 40% of the time for valence.

There are several things notable about this. First, proponents of error and bias do not have much to fear. These results are the same as saying people were *wrong* 78% and 60% of the time, respectively, for content and valence. Second, however, given the difficulty of the task—essentially, reading a complete stranger's thoughts and feelings in a 6-minute interaction—I consider these levels of accuracy (both of which were statistically significant) to be extraordinary.

Third, the pattern (not merely the overall levels) was actually different for content and valence. Content accuracy occurred because the participants succeeded at understanding what their partner, specifically, was thinking at a particular point. Valence accuracy, however, did not occur because people were good at reading their partner's specific feelings at a specific point. It occurred, instead, because participants had a good idea of how the other students in the study, in general, would feel in the interaction. This is, in essence, a phenomenon related to types of accuracy discussed in prior chapters of this book. First, it is in the family of Cronbach's and Kenny's "elevation" effects discussed in Chapter 12, on componential approaches to accuracy. Second, it can be considered a type of stereotype accuracy (in the Chapters 16 and 17 sense), where the stereotype involves "the other college students in this experiment."

Stinson and Ickes (1992) followed up this research by testing the hypothesis that male friends would show greater empathic content accuracy than haphazard pairs of male strangers. They found that male friends were right about 36% of the time, male strangers about 24% of the time. Male friends were accurate 50% more often than were male strangers. These results, like the prior work, show more evidence of inaccuracy than of accuracy. Even friends were wrong more often than they were right. That friends were right considerably more often than strangers, however, further attests to the validity of the phenomenon of empathic accuracy and provides some preliminary evidence that, as people become closer friends, they actually do come to know each other better (although it is also possible that people who find it easier to read one another's thoughts and feelings are more likely to become friends in the first place). Since that time, research on empathic accuracy has accelerated, including studies of friends, romantic partners, and clinical contexts (there is too much such work to summarize here, but see Ickes, 1997, for an edited volume compiling such research).

ACCURACY AT "ZERO ACQUAINTANCE" AND FROM "THIN SLICES"

"Zero acquaintance" is the psychological jargony term that refers to judgments regarding strangers (i.e., they have never met before, so they are unacquainted, thus the term "zero acquaintance"). Often, this research involves judgments about people one does not even meet (e.g., judgments about people in photos or on videotapes). "Thin slices" refers to judging someone on the basis of some relatively small amount of information about them, where small amount sometimes refers to a very brief exposure (often less than a few minutes and sometimes as brief as a few seconds) or, sometimes, to information that is very sketchy and degraded in some way.

One particularly striking line of research shows that people can accurately judge others' sexual orientation on the basis of minimal information (Ambady, Hallahan, & Conner, 1999). They first had people rate targets' sexual orientation from either a silent 10-second

video clip of the person, a 1-second clip, or a still photo. Those ratings correlated with targets' self-reported sexual orientation over 0.80 for the 10-second clips, over 0.50 for the 1-second clips, and over 0.30 for the still photos. People were not perfect, but these are extraordinary levels of accuracy for judgments of a seemingly nonobservable characteristic with such minimal information.

These results are also relevant to stereotype accuracy. Apparently, even though one cannot see someone's sexual orientation, people are often quite accurate at linking the right physical, appearance, and behavior cues associated with sexual orientation to particular individuals.

The thin slices research, however, is not limited to sexual orientation. For example, in their review of the literature up to that time, Ambady and Rosenthal (1992) found the following, on the basis of very brief periods of observation (less than 5 minutes):

1. Observers predict clinical psychological treatment outcomes at better than chance levels.
2. Observers accurately infer teacher expectations for students.
3. Observers accurately predict student evaluations of a teacher.

This is only a small subset of the domains that Ambady and Rosenthal's (1992) review addressed. Others included detection of deception, voting behavior, and depression. The overall accuracy of predictions based on thin slices was $r = .39$, which would appear to be a stunning level of accuracy (a mere 0.01 below a bull's eye, based on the standards developed in Chapter 16), given the minimal information upon which judgments were based.

There was, however, good news for proponents of error and bias in their review. Apparently, judgments based on "thick slices"—that is, when observers had much more information—were generally not much better than those based on thin slices.

This pattern is actually inconsistent with that found by researchers studying empathic accuracy (who, you may remember, found that friends were more accurate in judging one another's thoughts than were strangers). This apparent conflict, then, of whether knowing someone longer increases accuracy or not, will have to await further research for resolution.

Notes

1. For the statistically inclined, this occurred because Steele and Aronson (1995) used analysis of covariance (ANCOVA) and reported *adjusted means*, not the actual means. This worked as follows. First, there were mean SAT score differences between the African American and White students *prior* to the study. They used ANCOVA to control for these preexisting differences. So far, so good; this is completely reasonable and justified. ANCOVA can provide *statistically adjusted* means. This is essentially a prediction regarding what the mean scores on the outcome (SAT-like questions) would or should be in the two groups (stereotype threat vs. no stereotype threat) if there was no preexisting SAT difference. This, too, is completely justified, but it is very important to keep in mind what these adjusted means do and do not show. Their graph showed that, although there was no difference in outcome scores between African Americans and Whites in the nonstereotype threat condition, there was a substantial one in the stereotype threat

condition. Because, however, these were adjusted (not raw) means, what this actually means is the following:

- 1) The preexisting racial difference in SAT scores was unchanged in the nonstereotype threat conditions.
- 2) Stereotype threat actually increased the race difference in SAT scores.

The appearance of the graph, especially when read and interpreted by the statistically disinclined, however, visually depicted a very different and completely unjustified conclusion: That “but for stereotype threat” African American and White SAT scores would be the same. Now, Steele and Aronson (1995) did not actually draw that conclusion, but they did present the graph that made such a conclusion appear to be justified. Furthermore, there was no text making clear that what they found is described by points 1 and 2 above, which is why I think they do bear some responsibility for others drawing this extreme and unjustified conclusion. And, actually, it was not merely the statistically disinclined who misinterpreted their conclusions; some very statistically sophisticated people did so, too (see Sackett et al., 2004). When highly skilled scientists, trained to have a healthy scientific skepticism about almost everything, jettison that skepticism to embrace a finding like this, it is very tempting to think that there was something other than science going on.

2. It is most definitely possible to study religion or politics scientifically. For example, surveying a large random sample of some country’s population regarding their religious or political beliefs would be scientific. That is not my point here. My point is merely that it is often not possible to obtain scientific evidence regarding the validity of many religious or political beliefs.

*The most erroneous stories are those we think we know best - and
therefore never scrutinize or question.*

—STEPHEN JAY GOULD

21 The 90% Full Glass Contests the Bias for Bias

“The Story”

Research on expectancies has been interpreted by many scholars as providing a powerful and profound insight into a major source of social, educational, and economic inequality. Teacher expectations seemed to systematically advantage students from already advantaged backgrounds (e.g., Whites, middle class students, etc.) and disadvantage students from already disadvantaged backgrounds (e.g., ethnic minorities, students from lower social class backgrounds). To the extent that education is a major stepping stone toward occupational and economic advancement, self-fulfilling prophecies, it would seem, constituted a major social force operating to keep the oppressed in “their place.”

Fuel was further added to the fire of this sort of story by additional early research showing that social stereotypes can indeed be self-fulfilling (see Chapter 4). When men interviewed a woman who they falsely believed was physically attractive (accomplished through the use of false photographs and non-face-to-face interviews), not only were the men warmer and friendlier to her but also she became warmer and friendlier in response. When White interviewees were treated in the same cold and distant manner that White interviewers treated African Americans interviewees, the performance of the White interviewees suffered.

On the basis of these types of findings, and especially when combined with the assumption of stereotype inaccuracy, some scholars have concluded that self-fulfilling prophecies are likely to be a powerful and pervasive source of social injustice and group inequalities (e.g., Claire & Fiske, 1998; Jones, 1990; Weinstein, Gregory, & Strambler, 2004). Stereotypes lead to inaccurate expectations for individuals. Inaccurate expectations are powerfully and pervasively self-fulfilling. Because stereotypes are, the story goes, so widely shared and so widely inaccurate, their powerfully self-fulfilling effects will accumulate over time and across perceivers.

Because self-fulfilling prophecies are so consistently harmful to members of historically stigmatized groups, damaging self-fulfilling prophecy on top of damaging self-fulfilling prophecy will be heaped upon the backs of those already most heavily burdened by disadvantage and oppression. Thus, the achievement and advancement of people from stigmatized groups will be so repeatedly undermined by self-fulfilling prophecies that self-fulfilling prophecies constitute a major source of social inequalities and social problems.

The Inadequacy of “The Story” ---

The most benevolent interpretation of this sort of conclusion is that it is woefully incomplete. Cognitive biases do sometimes lead to expectancy confirmation and expectancies do sometimes lead to self-fulfilling prophecies. No doubt about it. In this sense, those telling this story are not completely wrong. But the power of expectations to distort social beliefs through biases and to create actual social reality through self-fulfilling prophecies is, in general, so small, fragile, and fleeting that it is quite difficult to make a convincing case based on a complete and careful reading of the actual scientific data that such effects likely constitute a major source of inequality. At minimum, those making this case, in their narrative reviews, have almost never grappled with the fact that hundreds of studies show that the biasing effects of expectations and stereotypes on person perception hover barely above zero (see Table 6–1 and Chapter 18), that self-fulfilling prophecy effects are often modest and fleeting (dissipating rather than increasing over time), and that some of the largest self-fulfilling prophecy effects ever obtained increased rather than decreased the performance of low-achieving students (see Chapters 3, 6, 13, and 14).

But a harsher view of “the story” may also be taken, one which concludes that “the story” is out of touch with the data available in 2012. It is either wrong in its particulars (depending on the particular claim) or so systematically distorts and overstates the evidence regarding the power of expectancies and stereotypes that it is fundamentally not credible. Abundant evidence attests to the implausibility of “the story”; very little actually supports it. “The story” is maintained primarily by a very selective and uncritical consideration of the evidence (see, e.g., almost any prior chapter in this book).

And what about accuracy? Except to be dismissed, it remains largely ignored when articles discuss the role of stereotypes or self-fulfilling prophecies in social problems. Despite the fact that stereotype accuracy is one of the largest effects in all of social psychology, social psychology textbooks spend pages and pages on bias and rarely mention accuracy. Similarly, far more space is typically provided to a handful of dramatic studies demonstrating self-fulfilling prophecies and biases than to the overwhelming evidence that accuracy is typically the single biggest source of “expectancy confirmation.” That is, social interaction *does* often confirm people’s expectations—because those expectations are often at least moderately accurate.

Why Is “The Story” So Popular? ---

One might wonder, then, why claims touting the power of stereotypes and expectancies have so dominated the social sciences. Well, this just might be one of those situations where the perspectives emphasizing the power of bias are right. Most social psychologists either really

want to believe in the power of bias or have been so steeped in the field's traditional emphasis on bias that it is part of the intellectual air. Claims to the contrary seem radical and unjustified. Thus, a truly powerful bias may have been born: the social psychological bias in favor of bias.

The process by which this may work may be quite subtle and invidious, despite the best of intentions of its proponents. I do not know how many times psychological researchers have said something to me along the lines of "I do not deny accuracy; I just find error and biases to be so much more interesting (or important)." That is fine. People are certainly allowed to have their own personal tastes in research topics as much as in anything else.

But there is a problem. When the overwhelming majority of the field considers error and bias to be more important and interesting, we end up with a scholarship that overwhelmingly investigates and demonstrates error and bias. This happens, not because laypeople's beliefs are so endlessly dominated by error and bias, but because psychological researchers have an endless "taste" for error and bias research. Once this state of affairs is established, however, it risks becoming self-sustaining and self-justifying (almost exactly as Allport [1955] predicted over 50 years ago; see Chapter 2). That is, new, innocent, even-handed scholars enter the field and are confronted with a nearly endless scholarship demonstrating bias after bias and error after error. So, when they write review articles, book chapters, or even the narrative sections of empirical articles, they will, with earnest sincerity, make claims like "the literature overwhelmingly demonstrates that laypeople are subject to a wide array of biases and distortions in judgment and, although demonstrations of such errors and distortions are common, there is little evidence of accuracy or rationality" (Chapters 5 and 15 present dozens of such quotes).

Such claims are 100% true—the scientific literature does demonstrate bias after bias and provides relatively little evidence of accuracy. This occurs, however, not because laypeople are overwhelmed by bias and are so rarely rational or accurate, but because *psychologists' interests lead them to perform studies of bias far more than studies of accuracy, to interpret their studies as more consistent with bias than they really are, and to support publishing studies demonstrating bias in more influential journals, so that it becomes literally true that "the scientific literature is filled with demonstrations of bias."* Although literally true, such claims are readily misinterpreted as also meaning something that is completely ***unjustified***: that error, bias, and self-fulfilling prophecy dominate over accuracy and are the primary ways that people's interpersonal expectations and social stereotypes relate to social reality. In contrast, as a broad and general conclusion, just the opposite is true: Accuracy dominates and error, bias, and self-fulfilling prophecy are the relatively unusual exceptions.

Recognizing this does require seeing through the fog created by the relentless drumbeat of bias, but does not require a great deal of scientific expertise. Mostly, it requires common sense and a modicum of college-level mathematical and statistical literacy. Common sense: Hastorf & Cantril (1954) found opposing college partisans differed on a grand total of six, yes, count 'em six, penalty calls in an entire football game and this is presented as a paragon of subjectivity? Rosenhan (1973) had people request admission to mental institutions complaining of hallucinations, and the staff does not diagnose them as sane—and this is presented as triumphant evidence of the power of labels? Darley & Gross (1983) present evidence of social class bias in person perception, and, even though two attempts to replicate it have failed, we should just keep citing it anyway to support claims about the power of bias? Snyder & Swann

(1978b) found that, when given choices only among leading questions, people chose questions to which answers would confirm their expectations, and, even though all subsequent research has shown that, when given the freedom to ask nonleading, diagnostic questions they overwhelmingly do so, we should keep citing Snyder & Swann (1978b) anyway as demonstrating that people seek to confirm their expectations?

Minimal mathematical and statistical literacy: The effect size in Rosenthal & Jacobson's (1968a,b) "dramatic" study of teacher expectations was $r=.15$, a figure that closely corresponds to that found in most of the subsequent research. This is a "powerful" effect? Self-fulfilling prophecy effects generally get smaller, not larger, over time; how, exactly, is this a basis for claiming they accumulate? Meta-analysis show stereotype effects are typically about $r=.10$, and individuating information effects are about $r=.70$. This is the justification for considering stereotypes a difficult-to-override "default" basis of person perception? Stereotype accuracy relationships are typically around .5 to .8, which puts them among the largest relationships in all of social psychology—this is justification for emphasizing stereotypes' inaccuracy and irrationality?

In addition to the bias for bias, "the story" may also be sustained by its political appeal as a basis for fighting oppression (see Chapters 3, 10, and 15 through 18). Fighting oppression is a good thing. But here again, we are faced with an apparent choice between laudatory goals. Sometimes social scientists' desire to contribute to a more fair and just society may appear to conflict with the results of their research. If stereotypes cannot be credibly condemned as massively invalid distortions; if expectancies do not bias perception, memory, and information-seeking to any great extent; and if self-fulfilling prophecies do not accumulate to create ever-increasing differences between demographic groups, it would seem that we have lost some valuable rhetorical tools for fighting oppression. We can no longer point fingers at laypeople's invalid social beliefs as major constructors of social reality and as major contributors to many social problems.

To me, however, this supposed conflict between fighting the good fight and reporting the results of our research in a fair and well-justified manner is more apparent than real, for several reasons. First, distorting or exaggerating our findings to achieve political ends ultimately undermines the credibility of the social sciences. Eventually, the truth usually outs. If people are much higher wattage than social scientists usually give them credit for, even if social scientists insist on purveying distortions ("self-fulfilling prophecies are powerful and pervasive") or illusions ("stereotypes are inherently inaccurate"),¹ intelligent laypeople will often come around to (justifiably) suspecting there is something wrong with the "science" (and social scientists will continue to bemoan the fact that their work is often ignored!).

There is, however, a second and even more damning reason that allowing our political goals to (dis)color our research conclusions damages the credibility of the social sciences. It leaves the social sciences wide open to (what would then be) the valid criticism that it is producing little more than politics masquerading as science. Pundits and laypeople would then feel completely justified in dismissing the social sciences as hopelessly politicized, and politicians who disagree with the politics would then have more than a little justification for ignoring it and cutting funding to support it.

This is not an argument for doing "pure" science (if that exists) and ignoring policy implications. When the science is performed and, especially, interpreted in a manner relatively free of blatant political agendas, when we make earnest efforts to reach conclusions based on

the data rather than interpret the data on the basis of our political preferences, and when those data have policy implications, the social sciences can and should inform policy. Problems arise when we allow our political preferences to excessively taint and distort our conclusions. One can usually tell this sort of thing is happening when a person (or a field) insists on maintaining a belief (e.g., “stereotypes are inaccurate”) in the face of overwhelming evidence against it.

There is, however, another entirely different set of more substantive (less political) reasons why we, as social scientists, should want to work hard to ensure that our politics do not distort our science. The very social problems that engage social scientists’ political concerns are far more likely to be solved by acknowledging than by denying the data. If we think we are curing a social ill by treating the wrong problem, we are likely to create a new problem and not solve the original one. This is obvious in medicine. If we treat a patient for cancer but the patient has Lyme disease, that patient is not likely to improve, and may get worse. If we mistakenly place the blame for some social problems on “inaccurate” social stereotypes and then spend time and resources trying to correct them, if many stereotypes are not inaccurate, our “cure” is not likely to have much effect. To the extent that scientific time and energy, institutional policies, and economic resources are directed toward minor or nonexistent sources of social problems, the actual sources of social problems will get less attention than they deserve. In this manner, the goodness of the intentions of those railing against inaccurate stereotypes is actually an obstruction to adopting constructive and effective policies for creating greater equality of opportunity.

Some political issues are questions of fact, even if answers are not yet known with certainty (e.g., How much is human activity causing global warming? Does having a demographically mixed classroom improve the academic achievement of all students in that classroom?). Many, however, are not, which gets to the final reason why we should strive to keep our politics out of our conclusions. Even though data can bear on how we advocate, to the extent that political positions are, essentially, based on morals, data are largely irrelevant. And it is usually very easy to tell whether one’s position is fundamentally moral or scientific. If scientific data could lead one to change one’s position, one’s policy position is based on science. For example, if you could see yourself becoming an opponent of diversity programs because scientific evidence showed that such programs do more harm than good, then your position on this issue is scientific.

If, on the other hand, no data could lead you to change your position, then your position is not scientific. Continuing with the same example, perhaps your commitment to egalitarianism is so strong that no social science data could convince you that diversity programs are dysfunctional. (Note: I am not claiming that they are actually dysfunctional; I am merely taking a policy conclusion [“diversity programs are good”] and pointing out the difference between basing that conclusion on politics/morals vs. science; scientific beliefs can be changed by data, whereas moral beliefs are rarely subject to evidence-based disconfirmation.) It is completely appropriate for people’s morals to inform or even determine their political attitudes and policy positions. What is not appropriate, however, is for that to be the case and then to pretend that one’s position is based on science.

To me, that is the litmus test for determining whether any particular belief is scientific or nonscientific (moral, religious, political, philosophical, etc.). Is it possible for data to convince you to change your mind? If so, then your belief is scientific; otherwise, it is not.

I am not saying anyone should change their mind much in response to a single piece of data or a single study. It might be 80 degrees today in parts of Alaska, but that is not going to convince me that Alaska is usually a warm place. However, if, over the next few years, the daily average temperature in Anchorage is 80 degrees, I will indeed change my view of Anchorage. Data can change my mind about the temperature of Alaska. And, even if you disagree with many of the conclusions I have reached throughout this book, if you can imagine data that would change your mind, then we have a respectful scientific disagreement. Perhaps you know of research that seems to refute my conclusions. Perhaps you think my analysis and critique of the existing research suffer from imperfections. That is all fine. Reasonable people may disagree.

But, if you care about the issues addressed in this book and, especially, if you disagree with the general themes of this book, I do ask you to take a moment and ask yourself: What evidence could convince you to change your mind? I can tell you what would change my mind. If the next 100 studies on interpersonal expectancies showed that self-fulfilling prophecies are typically much larger than accuracy, I would no longer claim the glass is 90% full. If the next 100 studies of stereotype accuracy showed that the average correlation between a belief about a group and that group's characteristics was below the .20 average in social psychology, I would no longer claim the glass is 90% full. If the next 100 studies of the role of stereotypes in person perception showed that the effect sizes for stereotype biases were .70 and reliance on relevant individuating information was .10, I would no longer claim that the glass is 90% full.

But those hundreds of studies do not exist. *Au contraire*. Given the overwhelming evidence of accuracy in many stereotypic beliefs, if you still believe that "stereotypes are inherently inaccurate," what would it take for you to change your mind? Given the overwhelming evidence that expectancies produce modest biases and self-fulfilling prophecies, if you still believe in the power of expectancy effects, what would it take for you to change your mind? Given the overwhelming evidence that people judge others far more on the basis of relevant individuating information (when available) than on stereotypes, what would it take for you to change your belief that stereotypes are a "default" and powerful basis for person perception? If you cannot answer these questions, we do not have a scientific disagreement. And, just as a wealth of scientific evidence supporting evolution is unlikely to change the views of a Creationist, there is no reason for anything in this book to influence your views whatsoever.

The purpose of this book has not been to get you to change your morals, religion, political ideology, or philosophy. My goals have been much more modest: to get you to consider the possibilities that social beliefs are often far more accurate than they are usually given credit for being, that biases and self-fulfilling prophecies are usually far weaker and more fleeting than they are usually given credit for being, and that, in general, people are much higher wattage than the social sciences usually credit them as being. Not according to moral philosophy, but according to data.

Toward a Balanced Social Science: The Role of Data

Are we going to be storytellers, selectively choosing dramatic studies (no matter how flawed, limited, or irreplicable) around which we can tell dramatic stories about the constructive

power of distorted social beliefs? Or, are we going to be scientists, who reach conclusions on the basis of the data from our accumulated collection of research—mountains of research, in some cases, such as the extraordinarily limited extent to which expectancies bias judgments and of the only somewhat less limited extent to which expectancies create self-fulfilling prophecies, and smaller amounts that bear on issues such as accumulation of self-fulfilling prophecies and whether they are generally helpful or harmful? Are we going to conclude that people are largely inaccurate on the basis of studies of bias that do not assess accuracy because stories about bias are somehow more compelling than stories about accuracy? Or are we going to actually assess the accuracy of social perception? Are we going to throw up our metaphorical hands in despair at assessing what people do most of the time, as did Fiske and Neuberg (1990, p. 21):

Note that one cannot ultimately resolve the overall question of whether category-based or individuating processes are more common in daily life. One cannot feasibly do a representative sample survey of people's impression formation processes: one can merely demonstrate the category-based and individuating processes occur under specifiable circumstances and then argue that those circumstances are more or less representative of life outside the lab. Assertions about actual frequencies of each process are simply not provable. . . . Our own position is that under ordinary conditions, people simply do not pay enough attention to individuate each other.

Unpacking this set of claims is instructive. First, most of it is true in a narrow literal sense. One cannot "ultimately" resolve this issue. But, then, science almost *never* "ultimately" resolves any issue. All that science does is attempt to provide sufficient evidence supporting some claim so that it becomes difficult to believe otherwise. If it provides enough such evidence from a sufficiently wide variety of sources over a sufficiently long period of time, it may become ridiculous to believe otherwise. However, even that claim is contingent on *further research*. Science is hypothetically open to all sorts of seemingly absurd claims, should the evidence be provided.

For example, evolutionary biology claims that billions of years ago life emerged from inanimate matter. Although comparable conditions might yield emergence of life anywhere and anytime, here on Earth, now, life only comes from life. Chairs do not turn into dolphins; lamps do not turn into snakes. However, if someday someone could provide replicable evidence of chairs turning into dolphins, evolutionary biology would have to change its theories regarding where life comes from. Until that time, however, it won't.

In that spirit, how can Fiske and Neuberg (1990; or anyone else!) dismiss the overwhelming evidence that stereotypes have minimal effects on person perception? The answer comes later in their quote: They claim that one can merely argue that the circumstances one studies in the lab apply outside of the lab. This liberates advocates of error and bias to make almost any claim, independent of the data. Why? Let us say 100 studies find no bias and 5 find bias. The situation is not quite that extreme (see Chapter 18), but that is not the point. Even if it were this extreme, all one need do is interpret those studies in such a manner that one concludes that the circumstances investigated in those 5 studies correspond more to real life than the circumstances studied in the 100 studies! And so, the claim that bias dominates could remain intact.

This is no metaphorical hypothetical. This is very close to the actual state of affairs. Hundreds and hundreds of studies, most performed before 1990, showed that stereotype effects were weak and easily eliminated, and that individuating information dominates person perception (see Chapters 5, 6, 17, and 18). And it is only by an extraordinarily selective reading of the literature that any other claim could be maintained.

One last aspect of their quote is worth noting. They claim that one cannot obtain evidence that bears on the frequency with which people rely on stereotypes versus individuating information. Instead, they claim, one can only argue that one's lab studies apply to the real world. Well, I suppose so, *if one only performs lab studies*. There is, however, an alternative. And that is to do some real-world research, outside the lab. When such research has been done, it usually provides the exact same pattern of results as the lab research—weak expectancy effects, weak stereotype effects, powerful individuating information effects, moderate to high accuracy (see Chapters 3, 13, 14, 17, and 18).

The foundation of science is data. One can speculate in the absence of data. Even when there are data, one can speculate that the existing data are “wrong” or “biased” in some way. But, until one obtains new data showing that the conclusions reached on the basis of the old data are wrong, one has no justifiable basis for declaring one's claims to be true. One can claim UFOs constitute alien visitations, and one might even be right. But, absent evidence of aliens per se, there is no scientific reason to support such a conclusion. One can claim stereotypes are powerfully biasing and self-fulfilling, but there is no scientific reason to support such a claim, at least not as a broad generalization. (I note here that I am not claiming powerful effects never happen or cannot happen. I am one of few social psychologists to have actually found any evidence of truly large self-fulfilling prophecies—Jussim, Eccles, & Madon, 1996; Madon, Jussim, & Eccles, 1997; and those found in the 1996 paper may be the largest ever found by any social psychologist. One dramatic study, however, does not justify a broad, general conclusion.)

I am sure that nothing in this book will change the minds of the many true believers in the power of stereotypes and expectancy-based biases. For the rest of you, though, I have one simple request. Don't believe me. Do with the accumulated social science data exactly what Fiske and Neuberg (1990) say you *should* do when judging a person. Just pay attention to the data. Not just your favorite data. All of the data. And if it is not possible to pay attention to all of the data (sometimes, there is just too much, or it requires too much professional expertise, etc.), at least avoid the pitfall of focusing your attention on the data that you want to be true. Instead, work hard to get the full, big picture.

Epilogue: The Election of Barack Obama and the 90% Full Glass

I was going to end with the above paragraph, but by the time I had gotten here, the United States had elected its first African American president. And how that happened so deliciously validates the big picture taken throughout this book that I could not resist the temptation to add this epilogue. If people's beliefs and stereotypes were such powerful influences on perception and judgment, if bias was so routinely powerful and pervasive, if America was so thoroughly the racist society its social science critics so often claim, and if people were routinely so far out to lunch, so low wattage as psychology depicts them, Obama could never have been elected.

A Prediction Based on "The Story"

Obama's election does not signal the end of prejudice. Prejudice is alive and well, and will probably stay that way for a very long time. Obama's election, however, does provide deep, profound evidence disconfirming "the story" about the power of expectancies, stereotypes, prejudice, etc. No one has ever told this story in exactly this way. But by piecing together various aspects of the story from various places, one could easily tell it this way:

Unconscious racism is rampant in America (Chen & Bargh, 1997; Greenwald & Krieger, 2006; Kang & Banaji, 2006). Obama can hardly expect to avoid being viewed through the distorting power of stereotypes to bias and twist perception and judgment (Darley & Gross, 1983; Devine, 1995; Fiske, 1998; Fiske & Neuberg, 1990). Furthermore, the well-established (*sic*) tendencies for expectancies to direct attention (Jones, 1986) and lead people to seek expectancy-confirming evidence (Snyder & Swann, 1978b) will all but ensure that, even if Obama has a stereotype-disconfirming message, many people will not get it. Instead, they will selectively seek out and focus on information that confirms their prior expectations. The research on prejudice, stereotypes, and expectancies, therefore, predicts that the obstacles to electing an African American president are likely to be prohibitively large.

It's a good story, right? It sounds good, it is internally consistent, and it flows well. It clearly, however, has a problem. The prediction based on this story has been disconfirmed by the data. Funny thing, data.

The Other Story

Now, one could tell a very different and (perhaps to some) far less righteously satisfying story:

Although stereotypes, prejudice, and discrimination exist, most people, most of the time, judge individuals on their merits, that is, on their individuating information. There are many situations where people do not have much individuating information, and in these situations, they undoubtedly act on their prejudices and stereotypes in discriminatory ways. But the U.S. presidential elections, including the primaries, provided more than ample opportunity for people to get to know their candidates. Although there may be some people who will or will not vote for a candidate primarily on the basis of race, that number is likely to be so small that, at minimum, a strong minority candidate would have as fair a chance as anyone else of being elected president.

I cannot say I knew Obama would win when he first announced his candidacy. But I did believe he had an excellent chance, on several grounds:

1. I thought his early stance opposing the (by then highly unpopular) Iraq war would win supporters;

2. I thought he had an unusual ability to articulate² a clear and inspiring vision that might sway many Americans; and, most relevant to this book,
3. I believed the data testifying to the power of individuating information. I believed that many Americans, even those with prejudice in their hearts and stereotypes in their heads (even inaccurate ones), would be receptive to his message and could be swayed by what seemed to be candidate Obama's strengths: his judgment, his policy positions, and his ability to cross various divides (Black/White, Democrat/Republican, liberal/conservative, etc. — whether his Presidency has lived up to his apparent strengths as a candidate is beyond the scope of this book).

Although the election of Obama does not signal the end of prejudice, it does demonstrate the weakness of stereotypes in the face of clear and abundant individuating information.

The Extraordinarily Small Role of Obama's Race in the Election ("Bias Is Real but Small" Rides Again!)

Indeed, the role of Obama's race, per se, in the election turned out to be almost completely trivial. How can this be, when about 95% of Black voters chose Obama but only about 43% of White voters chose him? That looks like a big race effect, right? Well it is a big race difference among voters, but it is not much of an effect of Obama's race. Democratic presidential candidates typically receive about 90% of the Black vote (Observationalism, 2008). So, Obama received only a very slightly higher proportion of Black votes than did Kerry, Gore, etc. There was not much effect of Obama's race on the proportion of the Black vote.

What about the White vote? Obama received a *higher* proportion of the White vote than did Kerry or Gore (Observationalism, 2008). So, the overall numbers indicate that Obama's race was not very important. What about the social science data?

Data from a variety of sources converge on the conclusion that about 5% to 7% of the voters did not vote for Obama because he is Black. This is almost exactly what is predicted by the bias results shown in Table 6-1. This was the conclusion reached by national surveys conducted by Yahoo/Stanford University and Gallup before the election (Gallup, 2008; U.S. News, 2008). And it is the conclusion we reached in our own small-scale study conducted on Rutgers undergraduates during the primaries (Stevens, Cohen, & Jussim, 2008).

Of course, it is important to keep in mind what these numbers mean. If, say, 6% of the voters did not vote for Obama because of his race, that is the same thing as saying that 94% chose their candidate for reasons unrelated to Obama's race. People are not perfect. Bigotry is not dead. But that 94% number puts a smile on my face.

Of course, even that number may overstate the role of anti-Black racism in the election. First, I know of no research (yet) that examined the role of antiage prejudice in the election. Perhaps McCain lost as many or more votes because of his age as Obama did because of his race.

Second, what about the proportion of the vote that McCain lost because of his race? This is the type of question that almost never occurs to many social science researchers concerned about issues of racism, sexism, and bigotry. Fortunately, however, it did occur to the Gallup organization (Gallup, 2008), who asked voters whether they were more or less likely to vote

for Obama or McCain because of their race. Consistent with Table 6–1, Yahoo/Stanford, and my own study, they found that about 6% said they were less likely to vote for Obama because of his race. However, they also found that 9% said they were *more likely* to vote for him because of his race, and that 6% said they were less likely to vote for McCain because of *his race* (they also found that 7% said they were more likely to vote for McCain because of his race). But when you put all this stuff together, it appears as if:

1. racial preferences did play some small role in the election, and
2. there was little or no net disadvantage for Obama because he is Black.

The minimal role of race can also be seen another way. Neither McCain nor Obama was an incumbent president. Therefore, another way to evaluate the role of race in the election is to compare Obama's margin to that of other winners of presidential elections where there was no incumbent. If race played a major role, one would predict that his margin would be, on average, smaller than that of other winners in such elections. Of course, this is most likely to be true in an ideal world where "everything else was held constant," a condition that does not exist in the real world. Still, by going back far enough, one gets a sufficiently wide variety of situations that the comparison is worth making.

Since World War II, there have been six presidential elections wherein neither major party candidate was an incumbent president. Here are those elections (Leip, 2008), with the winner appearing in **bold font**, and the margin of victory to the right; further to the right are other notable events at the time.

1952, **Eisenhower** vs. Stevenson: 11%; highly unpopular war and sitting president
1960, Nixon vs. **Kennedy**: 0.27%; recession
1968, **Nixon** vs. Humphrey: 0.7%; highly unpopular war
1988, **G.H.W. Bush** vs. Dukakis: 8%.
2000, **G.W. Bush** vs. Gore: -0.49% (negative because Bush lost the popular vote)

Overall average margin of victory (1952–2000) = 3.9%. Obama's margin = 7%. If one considered only the Democratic margin, Obama does even better by comparison (on average, the Democratic candidate **lost** by almost 4%, so he did 11% better than the average Democratic candidate since World War II in elections without an incumbent president). So much for the prediction that Obama's margin of victory should be lower because of his race.

The world may never be perfected. Prejudice and bias will probably never be completely eradicated. But those results say something quite good about many American voters: high wattage, much higher than they are usually given credit for being by many social scientists. Not because they chose Obama per se: That is a matter of personal political preference. But, gentle reader, that is the point. People overwhelmingly made their choices on the basis of personal political preferences, and not primarily on the basis of the race of the candidates.

People are indeed subject to all sorts of imperfections, errors, and biases. And if one focuses only on those imperfections, one is likely to see a very empty glass. But, however true it may be that errors and biases exist, it is about time that the social sciences started

acknowledging that, with respect to social beliefs, social perception, and social reality, the big picture is that the glass is about 90% full.

Notes ---

1. I am not claiming that they necessarily do so intentionally. Even if unintentional, however, they are still distortions and illusions.

2. Will I be accused of being racist by characterizing Obama as able to “articulate” a strong vision? Some of you may remember the hullabaloo when, just before the Democratic primaries, one of the candidates described Obama as “articulate.” The idea was that this was some sort of underhanded racist slur, because it contrasted the well-spoken, inspiring, and eloquent Obama with either (depending on the person making the accusation) prior African American presidential candidates (e.g., Jesse Jackson, Al Sharpton) or perhaps even African Americans in general, who often speak with a more obvious dialect or accent. Remember who that candidate was, who so bigotedly referred to Obama as “articulate”? Joe Biden.

References

- Ahearn, L. (2001). Language and agency. *Annual Review of Anthropology*, 30, 109–137.
- Allen, B. P. (1995). Gender stereotypes are not accurate: A replication of Martin (1987) using diagnostic vs. self-report and behavioral criteria. *Sex roles*, 32, 583–600.
- Allport, F. H. (1955). *Theories of perception and the concept of structure*. New York: Wiley.
- Allport, G. W. (1954/1979). *The nature of prejudice* (2nd ed.). Cambridge, MA: Perseus Books.
- Altemeyer, B. (1981). *Right-wing authoritarianism*. Winnipeg, Canada: University of Manitoba Press.
- Alvidrez, J., & Weinstein, R. S. (1999). Early teacher perceptions and later student achievement. *Journal of Educational Psychology*, 91, 732–746.
- Alwin, D., & Hauser, R. M. (1975). The decomposition of effects in path analysis. *American Sociological Review*, 40, 37–47.
- Ambady, N., Hallahan, M., & Conner, B. (1999). Accuracy of judgments of sexual orientation from thin slices of behavior. *Journal of Personality and Social Psychology*, 77, 538–547.
- Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111, 256–274.
- American Association of University Women. (1992). *How schools shortchange girls*. Washington, DC: American Association of University Women Educational Foundation, The Wellesley College Center for Research on Women.
- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.
- American Psychological Association. (1991). In the Supreme Court of the United States: Price Waterhouse v. Ann B. Hopkins (Amicus curiae brief). *American Psychologist*, 46, 1061–1070.

- Anderson, S. M., & Bem, S. L. (1981). Sex typing and androgyny in dyadic interaction: Individual differences in responsiveness to physical attractiveness. *Journal of Personality and Social Psychology*, 41, 74-86.
- Archer, D., & Akert, R. M. (1977). Words and everything else: Verbal and nonverbal cues in social interpretation. *Journal of Personality and Social Psychology*, 35, 443-449.
- Aronson, E. (1999). *The social animal* (8th ed.). New York: Worth Publishers.
- Asch, S. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology*, 41, 258-290.
- Ashmore, R. D., & Del Boca, F. K. (1981). Conceptual approaches to stereotypes and stereotyping. In D. L. Hamilton (Ed.), *Cognitive processes in stereotyping and intergroup behavior* (pp. 1-35). Hillsdale, NJ: Erlbaum.
- Ashton, M. C., & Esses, V. M. (1999). Stereotype accuracy: Estimating the academic performance of ethnic groups. *Personality and Social Psychology Bulletin*, 25, 225-236.
- Babad, E., Inbar, J., & Rosenthal, R. (1982). Pygmalion, Galatea, and the Golem: Investigations of biased and unbiased teachers. *Journal of Educational Psychology*, 74, 459-474.
- Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist*, 54, 462-479.
- Bargh, J. A., & Thein, D. (1985). Individual construct accessibility, person memory, and the recall-judgment link: The case of information overload. *Journal of Personality and Social Psychology*, 49, 1129-1146.
- Baron, R. M., Albright, L., & Malloy, T. E. (1995). The effects of behavioral and social class information on social judgment. *Personality and Social Psychology Bulletin*, 21, 308-315.
- Baumeister, R. F. (1987). How the self became a problem: A psychological review of historical research. *Journal of Personality and Social Psychology*, 52, 163-176.
- Baumeister, R. F., & Bushman, B. J. (2007) *Social psychology and human nature*. Belmont, CA: Thomson Wadsworth.
- Bayton, J., McAllister, L., & Hamer, J. (1956). Race-class stereotypes. *The Journal of Negro Education*, 25, 75-78.
- Bellezza, F. S., & Bower, G. H. (1981). Person stereotypes and memory for people. *Journal of Personality and Social Psychology*, 41, 856-865.
- Bender, K. J. (1990). *Psychiatric medications: A guide for mental health professionals*. Newbury Park, CA: Sage.
- Berger, P. L., & Luckmann, T. (1966). *The social construction of reality: A treatise in the sociology of knowledge*. New York: Doubleday.
- Beyer, S. (1999). The accuracy of academic gender stereotypes. *Sex Roles*, 40, 787-813.
- Biernat, M. (1995). The shifting standards model: Implications of stereotype accuracy for social judgment. In Y. T. Lee, L. Jussim, & C. R. McCauley (Eds.), *Stereotype accuracy* (pp. 87-114). Washington, DC: American Psychological Association.
- Block, J. (1993). Studying personality the long way. In D. C. Funder, R. D. Parke, C. Tomlinson-Keasey, & K. Widaman (Eds.), *Studying lives through time: Personality and development* (pp. 9-41). Washington, DC: American Psychological Association.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley Interscience.
- Borgida, E., Rudman, L. A., & Manteufel, L. L. (1995). On the courtroom use and misuse of gender stereotyping research. *Journal of Social Issues*, 51, 181-192.
- Brannon, L. (1999). *Gender: Psychological perspectives*. Boston: Allyn & Bacon.

- Brattesani, K. A., Weinstein, R. S., & Marshall, H. H. (1984). Student perceptions of differential teacher treatment as moderators of teacher expectation effects. *Journal of Educational Psychology, 76*, 236–247.
- Brewer, M. B. (1988). A dual process model of impression formation. In T. K. Srull & R. S. Wyer, Jr. (Eds.), *Advances in social cognition* (Vol. 1, pp. 1–36). Hillsdale, NJ: Erlbaum.
- Brewer, M. B., Dull, V., & Lui, L. (1981). Perceptions of the elderly: Stereotypes as prototypes. *Journal of Personality and Social Psychology, 41*, 656–670.
- Brigham, J. C. (1971). Ethnic stereotypes. *Psychological Bulletin, 76*, 15–38.
- Briton, N. J., & Hall, J. A. (1995). Beliefs about female and male nonverbal communication. *Sex Roles, 32*, 79–90.
- Brodt, S. E., & Ross, L. D. (1998). The role of stereotyping in overconfident social prediction. *Social Cognition, 16*, 225–252.
- Brophy, J. (1983). Research on the self-fulfilling prophecy and teacher expectations. *Journal of Educational Psychology, 75*, 631–661.
- Brophy, J., & Good, T. (1974). *Teacher-student relationships: Causes and consequences*. New York: Holt, Rinehart, and Winston.
- Brunswik, E. (1952). *The conceptual framework of psychology*. Chicago: University of Chicago Press.
- Campbell, D. T. (1967). Stereotypes and the perception of group differences. *American Psychologist, 22*, 817–829.
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand-McNally.
- Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Beverly Hills, CA: Sage.
- Carnegie, D. (1936). *How to win friends and influence people*. New York: Simon & Schuster.
- Cejka, M. A., & Eagly, A. H. (1999). Gender-stereotypic images of occupations correspond to the sex segregation of employment. *Personality and Social Psychology Bulletin, 25*, 413–423.
- Cerulo, K. A. (1997). Identity construction: New issues, new directions. *Annual Review of Sociology, 23*, 385–409.
- Chapman, G. B., & McCauley, C. (1993). Early career achievements of National Science Foundation (NSF) Graduate Applicants: Looking for Pygmalion and Galatea effects on NSF winners. *Journal of Applied Psychology, 78*, 815–820.
- Chen, M., & Bargh, J. A. (1997). Nonconscious behavioral confirmation processes: The self-fulfilling consequences of automatic stereotype activation. *Journal of Experimental Social Psychology, 33*, 541–560.
- Clabaugh, A., & Morling, B. (2004). Stereotype accuracy of ballet and modern dancers. *Journal of Social Psychology, 144*, 31–48.
- Claire, T., & Fiske, S. (1998). A systemic view of behavioral confirmation: Counterpoint to the individualist view. In C. Sedikides, J. Schopler, & C. A. Insko (Eds.), *Intergroup cognition and intergroup behavior* (pp. 205–231). Mahwah, NJ: Erlbaum.
- Clark, L. F., & Woll, S. B. (1981). Stereotype biases: A reconstructive analysis of their role in reconstructive memory. *Journal of Personality and Social Psychology, 41*, 1064–1072.
- Clarke, R. B., & Campbell, D. T. (1955). A demonstration of bias in the estimates of Negro ability. *Journal of Abnormal and Social Psychology, 51*, 585–588.

- Cline, V. B. (1964). Interpersonal perception. In B. Maher (Ed.), *Progress in experimental personality research* (Vol. 1, pp. 221–284). New York: Academic Press.
- Cohen, C. E. (1981). Personal categories and social perception: Testing some boundaries of the processing effects of prior knowledge. *Journal of Personality and Social Psychology*, 40, 441–452.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.
- Coleman, L., Jussim, L., & Hatter-Kelly, S. (1995). The nature of stereotyping: Utilizing three theories in a sample of Blacks. *Journal of Black Psychology*, 21, 332–356.
- Coles, R. (1969, April). What can you expect? (Review of the book, Pygmalion in the classroom.) *The New Yorker*, pp. 169–170, 173–177.
- Cook, M. (1979). *Perceiving others: The psychology of interpersonal perception*. London: Methuen.
- Cook, S. (1984). Cooperative interaction in multiethnic contexts. In N. Miller & M. B. Brewer (Eds.), *Groups in contact* (pp. 155–185). Orlando, FL: Academic Press.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.
- Cooper, H. (1979). Pygmalion grows up: A model for teacher expectation communication, and performance influence. *Review of Educational Research*, 49, 389–410.
- Cooper, H., & Hazelrigg, P. (1988). Personality moderators of interpersonal expectancy effects: An integrative research review. *Journal of Personality and Social Psychology*, 55, 937–949.
- Copus, D. (2004). Strangers in paradise: Junk science and gender stereotyping. Unpublished manuscript.
- Crano, W. D., & Mellon, P. M. (1978). Causal influence of teachers' expectations on children's academic performance: A cross-lagged panel analysis. *Journal of Educational Psychology*, 70, 39–49.
- Crocker, J., Major, B., & Steele, C. (1998). Social stigma. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 2, pp. 504–553). New York: McGraw-Hill.
- Cronbach, L. J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity." *Psychological Bulletin*, 52, 177–193.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- Danziger, K. (1997). The historical formation of selves. In R. D. Ashmore & L. Jussim (Eds.), *Self and identity: Fundamental issues* (pp. 137–159). New York: Oxford University Press.
- Darley, J. M., & Fazio, R. H. (1980). Expectancy–confirmation processes arising in the social interaction sequence. *American Psychologist*, 35, 867–881.
- Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, 44, 20–33.
- Davison, H. K., & Burke, M. J. (2000). Sex discrimination in simulated employment contexts: A meta-analytic investigation. *Journal of Vocational Behavior*, 56, 225–248.
- Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34, 571–582.
- Dawes, R. M. (1988). *Rational choice in an uncertain world*. San Diego: Harcourt Brace Jovanovich.
- Dawes, R. M. (1994). Affirmative action programs: Discontinuities between thoughts about individuals and thoughts about groups. In L. Heath, et al., (Eds.), *Applications of heuristics and biases to social issues* (pp. 223–239). New York: Plenum.

- Dawes, R. M. (2001). *Everyday irrationality: How pseudo-scientists, lunatics, and the rest of us systematically fail to think rationally*. Boulder, CO: Westview Press.
- Dawes, R., Singer, D., & Lemons, F. (1972). An experimental analysis of the contrast effect and its implications for intergroup communication and the indirect assessment of attitude. *Journal of Personality and Social Psychology*, 21, 281–295.
- Deaux, K., & Emswiller, T. (1974). Explanations of successful performance on sex-linked tasks: What is skill for the male is luck for the female. *Journal of Personality and Social Psychology*, 29, 80–85.
- Deaux, K., & Major, B. (1987). Putting gender into context: An interactive model of gender-related behavior. *Psychological Review*, 94, 369–389.
- Detterman, D. K., & Thompson, L. A. (1997). What is so special about special education? *American Psychologist*, 52, 1082–1090.
- Devine, P. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, 56, 5–18.
- Devine, P. (1995). Prejudice and outgroup perception. In A. Tesser (Ed.), *Advanced social psychology* (pp. 467–524). New York: McGraw-Hill.
- Devine, P. G., Hirt, E. R., & Gehrke, E. M. (1990). Diagnostic and confirmation strategies in trait hypothesis testing. *Journal of Personality and Social Psychology*, 58, 952–963.
- Dickman, A. B., Eagly, A. H., & Kulesa, P. (2002). Accuracy and bias in stereotypes about the social and political attitudes of women and men. *Journal of Experimental Social Psychology*, 38, 268–282.
- Dovidio, J. F., & Gaertner, S. L. (2010). Intergroup bias. In S. T. Fiske, D. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (5th ed., Vol. 2, pp. 1084–1121). Hoboken, NJ: Wiley.
- Downey, G., & Feldman, S. I. (1996). Implications of rejection sensitivity for close relationships. *Journal of Personality and Social Psychology*, 70, 1327–1343.
- Downey, G., Freitas, A. L., Michaelis, B., & Khouri, H. (1998). The self-fulfilling prophecy in close relationships: Rejection sensitivity and rejection by romantic partners. *Journal of Personality & Social Psychology*, 75, 545–560.
- Doyle, W. J., Hancock, G., & Kifer, E. (1972). Teachers' perceptions: Do they make a difference? *Journal of the Association for the Study of Perception*, 7, 21–30.
- Duncan, B. L. (1976). Differential social perception and attributions of intergroup violence: Testing the lower limits of stereotyping of blacks. *Journal of Personality and Social Psychology*, 34, 590–598.
- Eagly, A. H., & Diekmann, A. B. (2005). What is the problem? Prejudice as an attitude-in-context. In J. F. Dovidio, P. Glick, and L. A. Rudman (Eds.), *On the nature of prejudice: Fifty years after Allport* (pp. 19–35). Maiden, MA: Blackwell.
- Eagly, A. H., Makhijani, M. G., Ashmore, R. D., & Longo, L. C. (1991). What is beautiful is good, but...: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin*, 110, 109–128.
- Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motivation*, (pp. 75–146). San Francisco, CA: W. H. Freeman.
- Eccles, J. S., & Harold, R. D. (1992). Gender differences in educational and occupational patterns among the gifted. In N. Colangelo, S. G. Assouline, & D. L. Ambrosio (Eds.), *Talent development: Proceedings from the 1991 Henry B. and Jocelyn Wallace National Research Symposium on Talent Development*, (pp. 3–29). Unionville, NY: Trillium Press.

- Eccles (Parsons), J. (1984). Sex differences in mathematics participation. In M. L. Maehr & M. W. Steinkamp (Eds.), *Women in science, Vol. 2, Advances in motivation and achievement* (pp. 93–137). Greenwich, CT: JAI Press.
- Eccles, J. S., Wigfield, A., Flanagan, C. A., Miller, C., Reuman, D., & Yee, D. (1989). Self-concepts, domain values, and self-esteem: Relations and changes at early adolescence. *Journal of Personality*, 57, 283–310.
- Eccles, J., & Wigfield, A. (1985). Teacher expectations and student motivation. In J. Dusek (Ed.), *Teacher expectancies* (pp. 185–226). Hillsdale, NJ: Erlbaum.
- Eden, D., & Shani, A. B. (1982). Pygmalion goes to boot camp: Expectancy, leadership, and trainee performance. *Journal of Applied Psychology*, 67, 194–199.
- Ekman, P., & Friesen, W. V. (1969). Nonverbal leakage and cues to deception. *Psychiatry*, 32, 88–105.
- Ekman, P., & Friesen, W. V. (1974). Detecting deception from the body or face. *Journal of Personality and Social Psychology*, 29, 288–298.
- Elashoff, J. D., & Snow, R. E. (1971). *Pygmalion reconsidered*. Worthington, OH: Charles A. Jones.
- Ellison, J. M. (Ed.). (1989). *The psychotherapist's guide to psychopharmacotherapy*. Chicago: Year Book Medical Publishers.
- Fazio, R. H., Effrein, E. A., & Falender, V. J. (1981). Self-perceptions following social interaction. *Journal of Personality and Social Psychology*, 41, 232–242.
- Fein, S., & Spencer, S. J. (1997). Prejudice as self-image maintenance: Affirming the self through derogating others. *Journal of Personality and Social Psychology*, 73, 31–44.
- Feingold, A. (1992). Good-looking people are not what we think. *Psychological Bulletin*, 111, 304–341.
- Feldman, J. (1972). Stimulus characteristics and subject prejudice as determinants of stereotype attribution. *Journal of Personality and Social Psychology*, 21, 333–340.
- Felson, R. B. (1984). The effect of self-appraisals of ability on academic performance. *Journal of Personality and Social Psychology*, 47, 944–952.
- Festinger, L. (1957). *A theory of cognitive dissonance*. Palo Alto, CA: Stanford University Press.
- Finn, J. (1972). Expectations and the educational environment. *Review of Educational Research*, 42, 387–410.
- Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. In D. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 2, pp. 357–411). New York: McGraw-Hill.
- Fiske, S.T. (2004). *Social beings: A core motives approach to social psychology*. New York: Wiley.
- Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 1–74). New York: Academic Press.
- Fiske, S. T., & Taylor, S. E. (1984). *Social cognition*. Reading, MA: Addison-Wesley.
- Fiske, S. T., & Taylor, S. E. (1991). *Social cognition* (2nd ed.). New York: McGraw-Hill.
- Frieze, I. H., Olson, J. E., & Russell, J. (1991). Attractiveness and income for men and women in management. *Journal of Applied Social Psychology*, 21, 1039–1057.
- Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin*, 101, 75–90.

- Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, 102, 652–670.
- Gage, N. L., & Cronbach, L. J. (1955). Conceptual and methodological problems in interpersonal perception. *Psychological Review*, 62, 411–422.
- Gallup. (2008, October 9). *Obama's race may be as much a plus as a minus*. Retrieved December 8, 2008, from <http://www.gallup.com/poll/111049/Obamas-Race-May-Much-Plus-Minus.aspx>
- Gergen, K. J. (1985). The social constructionist movement in modern psychology. *American Psychologist*, 40, 266–275.
- Gersten, R., Walker, H. M., & Darch, C. (1988). Relationship between teachers' effectiveness and their tolerance for handicapped students: An exploratory study. *Exceptional Children*, 54, 433–438.
- Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton-Mifflin.
- Gilbert, D. T. (1995). Attribution and interpersonal perception. In A. Tesser (Ed.), *Advanced social psychology* (pp. 99–147). New York: McGraw-Hill.
- Gilbert, D. T. (1998). Ordinary personology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th edition, pp. 89–150). New York: McGraw-Hill.
- Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reason in everyday life*. New York: The Free Press.
- Glick, P., & Fiske, S. T. (1996). The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, 70, 491–512.
- Goldberg, P. (1968). Are women prejudiced against women? *Transaction*, 5, 28–30.
- Goldman, W., & Lewis, P. (1977). Beautiful is good: Evidence that the physically attractive are more socially skillful. *Journal of Experimental Social Psychology*, 13, 125–130.
- Gosling, S. D., Ko, S. J., Mannarelli, T., & Morris, M. E. (2002). A room with a cue: Personality judgments based on offices and bedrooms. *Journal of Personality and Social Psychology*, 82, 379–398.
- Gottfredson, L. S. (1994). The science and politics of race-norming. *American Psychologist*, 49, 955–963.
- Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence*, 24, 13–23.
- Gould, S. J. (1978). *The mismeasure of man*. New York: Norton.
- Graber, M. A., Bergus, G., Dawson, J. D., Wood, G. B., Levy, B. T., & Levin, I. (2000). Effect of a patient's psychiatric history on physicians' estimation of probability of disease. *Journal of General Internal Medicine*, 15, 204–206.
- Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102, 4–27.
- Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, 109, 3–25.
- Greenwald, A. G., & Krieger, L. H. (2006). Implicit bias: Scientific foundations. *California Law Review*, 94, 945–967.
- Hall, J. A., & Carter, J. D. (1999). Gender-stereotype accuracy as an individual difference. *Journal of Personality and Social Psychology*, 77, 350–359.
- Hall, J. A., Coats, E. J., & LeBeau, L. S. (2005). Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. *Psychological Bulletin*, 131, 898–924.

- Hamilton, D. L., & Rose, T. L. (1980). Illusory correlation and the maintenance of social stereotypes. *Journal of Personality and Social Psychology*, 39, 832–845.
- Hamilton, D. L., Sherman, S. J., & Ruvalo, C. M. (1990). Stereotype-based expectancies: Effects on information processing and social behavior. *Journal of Social Issues*, 46, 35–60.
- Hare-Mustin, R. T., & Maracek, J. (1988). The meaning of difference: Gender theory, postmodernism, and psychology. *American Psychologist*, 43, 455–464.
- Harris, M. J. (1991). Controversy and cumulation: Meta-analysis and research on interpersonal expectancy effects. *Personality and Social Psychology Bulletin*, 17, 316–322.
- Harris, M. J., Milich, R., Corbitt, E. M., Hoover, D. W., & Brady, M. (1992). Self-fulfilling effects of stigmatizing information on children's social interactions. *Journal of Personality and Social Psychology*, 63, 41–50.
- Harris, M. J., & Rosenthal, R. (1985). Mediation of interpersonal expectancy effects: 31 meta-analyses. *Psychological Bulletin*, 97, 363–386.
- Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world*. Thousand Oaks, CA: Sage Publications.
- Hastie, R., & Kumar, P. (1979). Person memory: Personality traits as organizing principles in memory for behaviors. *Journal of Personality and Social Psychology*, 37, 25–38.
- Hastorf, A. H., & Cantril, H. (1954). They saw a game: A case study. *Journal of Abnormal and Social Psychology*, 47, 129–143.
- Heath, L., et al. (Eds.). (1994). *Applications of heuristics and biases to social issues*. New York: Plenum.
- Hedges, S. M., Jandorf, L., & Stone, A. A. (1985). Meaning of daily mood assessments. *Journal of Personality and Social Psychology*, 48, 428–434.
- Heine, S. J., Buchtel, E. E., & Norenzayan, A. (2008). *Psychological Science*, 19, 309–313.
- Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist*, 58, 78–79.
- Herek, G. M. (1993). Documenting prejudice against lesbians and gay men on campus: The Yale Sexual Orientation Survey. *Journal of Homosexuality*, 25, 15–30.
- Herek, G. M. (2000). The psychology of sexual prejudice. *Current Directions in Psychological Science*, 9, 19–22.
- Herrnstein, R. J., & Murray, C. (1994). *The bell curve*. New York: The Free Press.
- Hilton, J., & Darley, J. (1985). Constructing other persons: A limit on the effect. *Journal of Experimental Social Psychology*, 21, 1–18.
- Hilton, J. L., & Fein, S. (1989). The role of typical diagnosticity in stereotype-based judgments. *Journal of Personality and Social Psychology*, 57, 201–211.
- Hinnant, J. B., O'Brien, M., & Ghazarian, S. R. (2009). The longitudinal relations of teacher expectations to achievement in the early school years. *Journal of Educational Psychology*, 101, 662–670.
- Hofer, M. A. (1994, December 26). Behind the curve. *The New York Times*, pp. A39.
- Hoge, R. D., & Butcher, R. (1984). Analysis of teacher judgments of pupil achievement levels. *Journal of Educational Psychology*, 76, 777–781.
- Holland, D. (1997). Selves as cultured: As told by an anthropologist who lacks a soul. In R. D. Ashmore & L. Jussim (Eds.), *Self and identity: Fundamental Issues* (pp. 160–190). New York: Oxford University Press.
- Humphreys, L. G., & Stubbs, J. (1977). A longitudinal analysis of teacher expectation, student expectation, and student achievement. *Journal of Educational Measurement*, 14, 261–270.

- Ickes, W. (Ed.). (1997). *Empathic accuracy*. New York: Guilford Press.
- Ickes, W., Stinson, L., Bissonnette, V., & Garcia, S. (1990). Naturalistic social cognition: Empathic accuracy in mixed sex dyads. *Journal of Personality and Social Psychology*, 59, 730-742.
- Investor's Business Daily. (1996). *Guide to the markets*. New York: Wiley.
- Itskowitz, R., Abend, T., & Dmitrovsky, L. (1986). The relationship between teachers' self-concept and their tendency to refer students for psychological help. *School Psychology International*, 7, 116-122.
- Jacoby, R., & Glauber, N. (1995). *The bell curve debate*. New York: Times Books.
- Jensen, A. R. (1969). How much can we boost I.Q. and scholastic achievement? *Harvard Educational Review*, 39, 1-123.
- Jones, E. E. (1985). Major developments in social psychology during the past five decades. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (3rd ed., Vol. 1, pp. 47-107). New York: Random House.
- Jones, E. E. (1986). Interpreting interpersonal behavior: The effects of expectancies. *Science*, 234, 41-46.
- Jones, E. E. (1990). *Interpersonal perception*. New York: W. H. Freeman and Company.
- Jones, J. M. (1996). *Prejudice and racism* (2nd ed.). New York: McGraw-Hill.
- Joreskog, K. G., & Sorbom, D. (1983). *Lisrel VI user's guide*. Chicago: International Educational Services.
- Jost, J. T., & Banaji, M. R. (1994). The role of stereotyping in system-justification and the production of false consciousness. *British Journal of Social Psychology*, 33, 1-27.
- Jost, J. T., & Hamilton, D. (2005). Stereotypes in our culture. In J. F. Dovidio, P. Glick, and L. A. Rudman (Eds.), *On the nature of prejudice: Fifty years after Allport* (pp. 208-224). Maiden, MA: Blackwell.
- Jost, J. T., & Kruglanski, A. W. (2002). The estrangement of social constructionism and experimental social psychology: History of the rift and prospects for reconciliation. *Personality and Social Psychology Review*, 6, 168-187.
- Judd, C. M., & Park, B. (1993). Definition and assessment of accuracy in social stereotypes. *Psychological Review*, 100, 109-128.
- Judd, C. M., Ryan, C. S., & Park, B. (1991). Accuracy in the judgment of in-group and out-group variability. *Journal of Personality and Social Psychology*, 61, 366-379.
- Jussim, L. (1986). Self-fulfilling prophecies: A theoretical and integrative review. *Psychological Review*, 93, 429-445.
- Jussim, L. (1989). Teacher expectations: Self-fulfilling prophecies, perceptual biases, and accuracy. *Journal of Personality and Social Psychology*, 57, 469-480.
- Jussim, L. (1990). Social reality and social problems: The role of expectancies. *Journal of Social Issues*, 46, 9-34.
- Jussim, L. (1991). Social perception and social reality: A reflection-construction model. *Psychological Review*, 98, 54-73.
- Jussim, L. (2005). Accuracy: Criticisms, controversies, criteria, components, and cognitive processes. *Advances in Experimental Social Psychology*, 37, 1-93.
- Jussim, L., Cain, T., Crawford, J., Harber, K., & Cohen, F. (2009). The unbearable accuracy of stereotypes. In T. Nelson (Ed.), *Handbook of prejudice, stereotyping, and discrimination* (pp. 199-227). Hillsdale, NJ: Erlbaum.
- Jussim, L., Coleman, L., & Lerch, L. (1987). The nature of stereotypes: A comparison and integration of three theories. *Journal of Personality and Social Psychology*, 52, 536-546.

- Jussim, L., Coleman, L., & Nassau, S. (1987). The influence of self-esteem on perceptions of performance and feedback. *Social Psychology Quarterly*, 50, 95-99.
- Jussim, L., & Eccles, J. (1992). Teacher expectations II: Reflection and construction of student achievement. *Journal of Personality and Social Psychology*, 63, 947-961.
- Jussim, L., & Eccles, J. (1995). Naturalistic studies of interpersonal expectancies. *Review of Personality and Social Psychology*, 15, 74-108.
- Jussim, L., Eccles, J., & Madon, S. J. (1996). Social perception, social stereotypes, and teacher expectations: Accuracy and the quest for the powerful self-fulfilling prophecy. *Advances in Experimental Social Psychology*, 29, 281-388.
- Jussim, L., Fleming, C., Coleman, L., & Kohlberger, C. (1996). The nature of stereotypes II: A multiple-process model of evaluations. *Journal of Applied Social Psychology*, 26, 283-312.
- Jussim, L., & Harber, K. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns; resolved and unresolved controversies. *Personality and Social Psychology Review*, 9, 131-135.
- Jussim, L., Harber, K. D., Crawford, J. T., Cain, T. R., & Cohen, F. (2005). Social reality makes the social mind: Self-fulfilling prophecy, stereotypes, bias, and accuracy. *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems*, 6, 85-102.
- Jussim, L., McCauley, C. R., & Lee, Y. T. (1995). Why study stereotype accuracy and inaccuracy? In Y. T. Lee, L. Jussim, & C. R. McCauley (Eds.), *Stereotype accuracy: Toward appreciating group differences* (pp. 3-27). Washington, DC: American Psychological Association.
- Jussim, L., Smith, A., Madon, S., & Palumbo, P. (1998). Teacher expectations. In J. Brophy (Ed.), *Advances in research on teaching* (Vol. 7, pp. 1-48). Greenwich, CT: JAI Press.
- Jussim, L., Soffin, S., Brown, R., Ley, J., & Kohlhepp, K. (1992). Understanding reactions to performance feedback by integrating ideas from symbolic interactionism and cognitive evaluation theory. *Journal of Personality and Social Psychology*, 62, 402-421.
- Jussim, L., Yen, H., & Aiello, J. (1995). Self-consistency, self-enhancement, and accuracy in reactions to feedback. *Journal of Experimental Social Psychology*, 31, 322-356.
- Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 237-251.
- Kang, J., & Banaji, M. R. (2006). Fair measures: A behavioral realist revision of "affirmative action". *California Law Review*, 94, 1063-1118.
- Katz, D., & Braly, K. (1933). Racial stereotypes of one hundred college students. *Journal of Abnormal and Social Psychology*, 28, 280-290.
- Kelley, H. H. (1950). The warm-cold variable in first impressions of persons. *Journal of Personality*, 18, 431-439.
- Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska symposium on motivation* (Vol. 15, pp. 292-328). Lincoln: University of Nebraska Press.
- Kelley, H. H., & Stahelski, A. J. (1970). The social interaction basis of cooperators' and competitors' beliefs about others. *Journal of Personality and Social Psychology*, 16, 66-91.
- Kelly, G. A. (1955). *The psychology of personal constructs* (Vol. 1, A theory of personality). New York: W. W. Norton and Company.

- Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. New York: Guilford.
- Kenny, D. A., & Albright, L. (1987). Accuracy in interpersonal perception: A social relations analysis. *Psychological Bulletin*, 102, 390-402.
- Kenny, D. A., & DePaulo, B. M. (1993). Do people know how others view them? An empirical and theoretical account? *Psychological Bulletin*, 114, 145-161.
- Kenny, D. A., & La Voie, L. (1984). The social relations model. *Advances in Experimental Social Psychology*, 18, 141-182.
- Kenny, D. A., West, T. V., Malloy, T. E., & Albright, L. (2006). Componential analysis of interpersonal perception data. *Personality and Social Psychology Review*, 4, 282-294.
- Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94, 211-228.
- Krosnick, J. A. (1991). Response strategies for coping with cognitive demands of attitudes measures in surveys. *Applied Cognitive Psychology*, 5, 213-236.
- Krueger, J. (1996). Probabilistic national stereotypes. *European Journal of Social Psychology*, 26, 961-980.
- Krueger, J., & Funder, D. C. (2004). Towards a balanced social psychology: Causes, consequences, and cures for the problem-seeking approach to social behavior and cognition. *Behavioral and Brain Sciences*, 27, 313-327.
- Krueger, J., & Rothbart, M. (1988). Use of categorical and individuating information in making inferences about personality. *Journal of Personality and Social Psychology*, 55, 187-195.
- Kruglanski, A. (1989). The psychology of being "right": The problem of accuracy in social perception and cognition. *Psychological Bulletin*, 106, 395-409.
- Kuklinski, M. R., & Weinstein, R. S. (2001). Classroom developmental differences in a path model of teacher expectancy effects. *Child Development*, 72, 1554-1578.
- Kulik, J. A. (1983). Confirmatory attribution and the perpetuation of social beliefs. *Journal of Personality and Social Psychology*, 44, 1171-1181.
- Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin*, 127, 162-181.
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108, 480-498.
- Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review*, 103, 284-308.
- LaPiere, R. T. (1934). Attitudes versus actions. *Social Forces*, 13, 230-237.
- LaPiere, R. T. (1936). Type-rationalizations of group antipathy. *Social Forces*, 15, 232-237.
- Lareau, A. (1987). Social-class differences in family-school relationships: The importance of cultural capital. *Sociology of Education*, 60, 73-85.
- Lee, Y. T., Jussim, L., & McCauley, C. R. (Eds.). (1995). *Stereotype accuracy: Toward appreciating group differences*. Washington, DC: American Psychological Association.
- Lee, Y. T., & Ottati, V. (1993). Determinants of ingroup and outgroup perceptions of heterogeneity: An investigation of Sino-American stereotypes. *Journal of Cross-Cultural Psychology*, 24, 298-318.
- Lee, Y. T., & Ottati, V. (1995). Accuracy: A neglected component of stereotype research. In Y. T. Lee, L. Jussim, & C. R. McCauley (Eds.), *Stereotype accuracy: Toward appreciating group differences* (pp. 3-27). Washington, DC: American Psychological Association.

- Leip, D. (2008). *Dave Leip's atlas of U.S. presidential elections*. Retrieved December 8, 2008, from <http://www.uselectionatlas.org/RESULTS/compare.php?type=national&year=2000&f=0&off=0&elect=0>
- Lenski, G. E., & Leggett, J. C. (1960). Caste, class, and deference in the research interview. *American Journal of Sociology*, 65, 463–467.
- Levine, R., Chein, I., & Murphy, G. (1942). The relation of the intensity of a need to the amount of perceptual distortion: A preliminary report. *The Journal of Psychology*, 13, 283–293.
- Linville, P. W., Fischer, G. W., & Salovey, P. (1989). Perceived distributions of the characteristics of in-group and out-group members: Empirical evidence and a computer simulation. *Journal of Personality and Social Psychology*, 57, 165–188.
- Linville, P. W., & Jones, E. E. (1980). Polarized appraisal of out-group members. *Journal of Personality and Social Psychology*, 38, 689–703.
- Lippmann, W. (1922/1991). *Public opinion*. New Brunswick, NJ: Transaction Publishers.
- Locksley, A., Borgida, E., Brekke, N., & Hepburn, C. (1980). Sex stereotypes and social judgment. *Journal of Personality and Social Psychology*, 39, 821–831.
- Locksley, A., Hepburn, C., & Ortiz, V. (1982). Social stereotypes and judgments of individuals: An instance of the base-rate fallacy. *Journal of Experimental Social Psychology*, 18, 23–42.
- Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37, 2098–2109.
- MacKay, C. (1841/1932). *Extraordinary popular delusions and the madness of crowds*. London: The Noonday Press.
- Mackie, D. M., Hamilton, D. L., Susskind, J., & Roselli, F. (1996). Social psychological foundations of stereotype formation. In C. N. Macrae, C. Stangor, & M. Hewstone (Eds.), *Stereotypes and stereotyping* (pp. 41–78). New York: Guilford.
- Mackie, M. (1973). Arriving at “truth” by definition: The case of stereotype inaccuracy. *Social Problems*, 20, 431–447.
- Macrae, C. N., Milne, A. B., & Bodenhausen, G. V. (1994). Stereotypes as energy-saving devices: A peek inside the cognitive toolbox. *Journal of Personality and Social Psychology*, 66, 37–47.
- Macrae, C. N., Stangor, C., & Hewstone, M. (1996). *Stereotypes and stereotyping*. New York: Guilford.
- Madon, S. (1997). What do people believe about gay males? A study of stereotype content and strength. *Sex Roles*, 37, 663–685.
- Madon, S., Gyll, M., Spoth, R., & Willard, J. (2004). Self-fulfilling prophecies: The synergistic accumulation of parents’ beliefs on children’s drinking behavior. *Psychological Science*, 15, 837–845.
- Madon, S. J., Jussim, L., & Eccles, J. (1997). In search of the powerful self-fulfilling prophecy. *Journal of Personality and Social Psychology*, 72, 791–809.
- Madon, S. J., Jussim, L., Keiper, S., Eccles, J., Smith, A., & Palumbo, P. (1998). The accuracy and power of sex, social class and ethnic stereotypes: Naturalistic studies in person perception. *Personality and Social Psychology Bulletin*, 24, 1304–1318.
- Madon, S. J., Smith, A., Jussim, L., Russell, D. W., Eccles, J., Palumbo, P., & Walkiewicz, M. (2001). Am I as you see me or do you see me as I am: Self-fulfilling prophecies versus self-verification. *Personality and Social Psychology Bulletin*, 27, 1214–1224.

- Major, B., Cozzarelli, C., Testa, M., & McFarlin, D. B. (1988). Self-verification versus expectancy confirmation in social interaction: The impact of self-focus. *Personality and Social Psychology Bulletin*, 14, 346–359.
- Malkiel, B. G. (1973). *A random walk down Wall Street*. New York: Norton.
- Marger, M. N. (1994). *Race and ethnic relations* (3rd ed.). Belmont, CA: Wadsworth.
- Markus, H. (1977). Self-schemas and processing information about the self. *Journal of Personality and Social Psychology*, 35, 63–78.
- Markus, H., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, 98, 224–253.
- Markus, H., & Zajonc, R. B. (1985). The cognitive perspective in social psychology. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (3rd ed., Vol. 1, pp. 137–230). New York: Random House.
- Martel, R. F., Lane, D. M., & Emrich, C. (1996). Male–female differences: A computer simulation. *American Psychologist*, 51, 157–158.
- Martin, C. L. (1987). A ratio measure of sex stereotyping. *Journal of Personality and Social Psychology*, 52, 489–499.
- Martin, C. L., & Parker, S. (1995). Folk theories about sex and race differences. *Personality and Social Psychology Bulletin*, 21, 45–57.
- Mazella, R., & Feingold, A. (1994). The effects of physical attractiveness, race, socioeconomic status, and gender of defendants and victims on judgments of mock jurors: A meta-analysis. *Journal of Applied Social Psychology*, 24, 1315–1344.
- McArthur, L. Z., & Baron, R. M. (1983). Toward an ecological theory of social perception. *Psychological Review*, 90, 215–238.
- McCauley, C. R. (1995). Are stereotypes exaggerated? A sampling of racial, gender, academic, occupational, and political stereotypes. In Y. T. Lee, L. Jussim, & C. R. McCauley (Eds.), *Stereotype accuracy: Toward appreciating group differences* (pp. 215–243). Washington, DC: American Psychological Association.
- McCauley, C., Jussim, L., & Lee, Y. T. (1995). Stereotype accuracy: Toward appreciating group differences. In Y. T. Lee, L. Jussim, & C. R. McCauley (Eds.), *Stereotype accuracy: Toward appreciating group differences* (pp. 293–312). Washington, DC: American Psychological Association.
- McCauley, C., & Stitt, C. L. (1978). An individual and quantitative measure of stereotypes. *Journal of Personality and Social Psychology*, 36, 929–940.
- McCauley, C., Stitt, C. L., & Segal, M. (1980). Stereotyping: From prejudice to prediction. *Psychological Bulletin*, 87, 195–208.
- McCauley, C., & Thangavelu, K. (1991). Individual differences in sex stereotyping of occupations and personality traits. *Social Psychology Quarterly*, 54, 267–279.
- McCauley, C., Thangavelu, K., & Rozin, P. (1988). Sex stereotyping of occupations in relation to television representations and census facts. *Basic and Applied Social Psychology*, 9, 197–212.
- Mcginnies, E. (1949). Emotionality and perceptual defense. *Psychological Review*, 56, 244–251.
- McKirnan, D. J., Smith, C. E., & Hamayan, E. V. (1983). A sociolinguistic approach to the belief-similarity model of racial attitudes. *Journal of Experimental Social Psychology*, 19, 434–447.
- McNatt, D. B. (2000). Ancient Pygmalion joins contemporary management: A meta-analysis of the result. *Journal of Applied Psychology*, 85, 314–322.

- McNulty, S. E., & Swann, W. B., Jr. (1994). Identity negotiation in roommate relationships: The self as architect and consequence of social reality. *Journal of Personality and Social Psychology*, 67, 1012–1023.
- Mead, M. (1956). The cross-cultural approach to the study of personality. In J. L. McCary (Ed.), *Psychology of personality* (pp. 201–252). New York: Grove Press.
- Meece, J. L., Eccles-Parsons, J., Kaczala, C. M., Goff, S. E., & Futterman, R. (1982). Sex differences in math achievement: Towards a model of academic choice. *Psychological Bulletin*, 91, 321–348.
- Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, 1, 108–141.
- Meijer, C., & Foster, S. (1988). The effect of teacher self-efficacy on referral chance. *Journal of Special Education*, 22, 378–385.
- Merton, R. K. (1948). The self-fulfilling prophecy. *Antioch Review*, 8, 193–210.
- Midgley, C., Feldlaufer, H., & Eccles, J. S. (1989). Change in teacher efficacy and student self- and task-related beliefs in mathematics during the transition to junior high school. *Journal of Educational Psychology*, 81, 247–258.
- Milgram, S. (1974). *Obedience to authority*. New York: Harper & Row.
- Miller, D. T., & Turnbull, W. (1986). Expectancies and interpersonal processes. *Annual Review of Psychology*, 37, 233–256.
- Murray, S. L., Holmes, J. G., & Griffin, D. W. (1996). The self-fulfilling nature of positive illusions in romantic relationships: Love is not blind, but prescient. *Journal of Personality and Social Psychology*, 71, 1155–1180.
- Myers, D. G. (1987). *Social psychology* (2nd ed.). New York: McGraw-Hill.
- Myers, D. G. (1999). *Social psychology* (6th ed.). New York: McGraw-Hill.
- Myers, D. G. (2002). *Social psychology* (7th ed.). New York: McGraw-Hill.
- Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.
- Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., et al. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, 51, 77–101.
- Nelson, T. (2002). *The psychology of prejudice*. Boston: Allyn & Bacon.
- Neuberg, S. L. (1989). The goal of forming accurate impressions during social interactions: Attenuating the impact of negative expectancies. *Journal of Personality and Social Psychology*, 56, 374–386.
- Neuberg, S. L. (1994). Expectancy-confirmation processes in stereotype-tinged social encounters: The moderating role of social goals. In M. P. Zanna & J. M. Olson (Eds.), *The psychology of prejudice: The Ontario symposium* (Vol. 7, pp. 103–130). Hillsdale, NJ: Erlbaum.
- The New York Times*. (1994, July 13). Top Jersey court orders new plan for school funds. *The New York Times*, pp. A1, B6.
- Niemann, Y. F., & Maruyama, G. (2005). Inequities in higher education: Issues and promising practices in a world ambivalent about affirmative action. *Journal of Social Issues*, 61, 407–426.
- Nisbett, R. E., & Cohen, D. (1996). *Culture of honor: Psychology of violence in the south*. Boulder, CO: Westview Press.
- Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review*, 108, 291–310.
- Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.

- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231–259.
- Nisbett, R. E., Zukier, H., & Lemley, R. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology*, 13, 248–277.
- Norenzayan, A., & Nisbett, R. E. (2000). Culture and causal cognition. *Current Directions in Psychological Science*, 9, 132–135.
- Norton, M. I., Sommers, S. R., Vandello, J. A., & Darley, J. M. (2006). Mixed motives and racial bias: The impact of legitimate and illegitimate criteria on decision making. *Psychology, Public Policy, and the Law*, 12, 36–55.
- Oakes, J. (1985). *Keeping track: How schools structure inequality*. New Haven, CT: Yale University Press.
- Oakes, P. J., Haslam, S. A., & Turner, J. C. (1994). *Stereotyping and social reality*. Cambridge, MA: Blackwell.
- Observationalism. (2008, November 9). *Selected exit poll comparisons, 2000–2004–2008*. Retrieved December 8, 2008, from <http://observationalism.com/2008/11/09/selected-exit-poll-comparisons-2000-2004-2008>
- Olson, J. M., Roese, N. J., & Zanna, M. P. (1996). Expectancies. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 211–239). New York: Guilford Press.
- Ottati, V., & Lee, Y. T. (1995). Accuracy: A neglected component of stereotype research. In Y. T. Lee, L. Jussim, & C. R. McCauley (Eds.), *Stereotype accuracy: Toward appreciating group differences* (pp. 29–59). Washington, DC: APA Press.
- Palardy, J. M. (1969). What teachers believe—what children achieve. *Elementary School Journal*, 69, 370–374.
- Park, B., & Judd, C. M. (2005). Rethinking the link between categorization and prejudice within the social cognition perspective. *Personality and Social Psychology Review*, 9, 108–130.
- Parsons, J. E., Kaczala, C. M., & Meece, J. L. (1982). Socialization of achievement attitudes and beliefs: Classroom influences. *Child Development*, 53, 322–339.
- Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes*, (pp. 17–59). San Diego: Academic Press.
- Paulhus, D. L. (1998). Interpersonal and intrapsychic adaptiveness of trait self-enhancement: A mixed blessing? *Journal of Personality and Social Psychology*, 74, 1197–1208.
- Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 67–88). Hillsdale, NJ: Erlbaum.
- Pickering, M. (2001). *Stereotyping*. New York: Palgrave.
- Pinderhughes, E. (1989). *Understanding race, ethnicity, and power: The key to efficacy in clinical practice*. New York: The Free Press.
- Plous, S. (2003). *Understanding prejudice and discrimination*. New York: McGraw-Hill.
- Popper, K. R. (1959/1968). *The logic of scientific discovery*. New York: Harper & Row.
- Rahn, W. M. (1993). The role of partisan stereotypes in information processing about political candidates. *American Journal of Political Science*, 37, 472–496.

- Räikkönen, K., Matthews, K. A., Flory, J. S., Owens, J. F., & Gump, B. B. (1999). Effects of optimism, pessimism, and trait anxiety on ambulatory blood pressure and mood during everyday life. *Journal of Personality and Social Psychology*, 76, 104–113.
- Raudenbush, S. W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy inductions: A synthesis of findings from 18 experiments. *Journal of Educational Psychology*, 76, 85–97.
- Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 301–321). New York: Russell Sage Foundation.
- Redding, R. E. (2001). Sociopolitical diversity in psychology: The case for pluralism. *American Psychologist*, 56, 205–215.
- Rettew, D. C., Billman, D., & Davis, R. A. (1993). Inaccurate perceptions of the amount others stereotype: Estimates about stereotypes of one's own group and other groups. *Basic and Applied Social Psychology*, 14, 121–142.
- Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, 7, 331–363.
- Rist, R. (1970). Student social class and teacher expectations: The self-fulfilling prophecy in ghetto education. *Harvard Educational Review*, 40, 411–451.
- Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (Eds.). (1991). *Measures of personality and social psychological attitudes*. San Diego: Academic Press.
- Robinson, R. J., Keltner, D., Ward, A., & Ross, L. (1995). Actual versus assumed differences in construal: "Naïve realism" in intergroup perception and conflict. *Journal of Personality and Social Psychology*, 68, 404–417.
- Rogers, R. W., & Prentice-Dunn, S. (1981). Deindividuation and anger-mediated interracial aggression: Unmasking regressive racism. *Journal of Personality and Social Psychology*, 41, 63–73.
- Rokeach, M., & Mezei, L. (1966). Race and shared belief as factors in social choice. *Science*, 151, 167–172.
- Rosenberg, S. (1977). New approaches to the analysis of personal constructs in person perception. In A. W. Landfield (Ed.), *Nebraska symposium on motivation* (Vol. 24, pp. 179–242). Lincoln: University of Nebraska Press.
- Rosenthal, D. L. (1973). On being sane in insane places. *Science*, 179, 250–258.
- Rosenthal, R. (1974). *On the social psychology of the self-fulfilling prophecy: Further evidence for Pygmalion effects and their mediating mechanisms*. New York: MSS Modular.
- Rosenthal, R. (1984). *Meta-analytic procedures for social research* (2nd ed.). Newbury Park, CA: Sage.
- Rosenthal, R. (1985). From unconscious experimenter bias to teacher expectancy effects. In J. Dusek (Ed.), *Teacher expectancies* (pp. 37–65). Hillsdale, NJ: Erlbaum.
- Rosenthal, R. (1989, August). *Experimenter expectancy, covert communication, & meta-analytic methods*. Invited address at the 97th Annual Convention of the American Psychological Association, New Orleans, LA.
- Rosenthal, R. (1991). *Meta-analytic procedures for social research* (2nd ed.). Newbury Park, CA: Sage.
- Rosenthal, R. (1995). Critiquing Pygmalion: A 25-year perspective. *Current Directions in Psychological Research*, 4, 171–172.
- Rosenthal, R., & Fode, K. L. (1963a). Three experiments in experimenter bias. *Psychological Reports*, 12, 491–511.

- Rosenthal, R., & Fode, K. L. (1963b). The effects of experimenter bias on the performance of the albino rat. *Behavioral Science*, 8, 183–189.
- Rosenthal, R., & Jacobson, L. (1968a). *Pygmalion in the classroom: Teacher expectations and student intellectual development*. New York: Holt, Rinehart, and Winston.
- Rosenthal, R., & Jacobson, L. F. (1968b). Teacher expectations for the disadvantaged. *Scientific American*, 218, 19–23.
- Rosenthal, R., & Lawson, R. (1964). A longitudinal study of the effects of experimenter bias on the operant learning of laboratory rats. *Journal of Psychiatric Research*, 2, 61–72.
- Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *The Behavioral and Brain Sciences*, 3, 377–386.
- Ross, L. D., Amabile, T. M., & Steinmetz, J. L. (1977). Social roles, social control, and biases in social-perception processes. *Journal of Personality & Social Psychology*, 35, 485–494.
- Ross, L. D., Lepper, M., & Ward, A. (2010). History of social psychology: Insights, challenges, and contributions to theory and application. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (5th ed., Vol. 1, pp. 3–50). Hoboken, NJ: John Wiley and Sons.
- Ross, L., & Nisbett, R. E. (1991). *The person and the situation*. New York: McGraw-Hill.
- Roth, B. M. (1995, January 2). We can throw teacher expectations on the IQ scrap heap. *The New York Times*, p. A25.
- Rothbart, M., Evans, M., & Fulero, S. (1979). Recall for confirming events: Memory processes and the maintenance of social stereotypes. *Journal of Experimental Social Psychology*, 15, 343–355.
- Rowe, D. C. (1995, January 2). Intervention fables. *The New York Times*, p. A25.
- Rubin, J. Z., Kim, S. H., & Peretz, N. M. (1990). Expectancy effects and negotiation. *Journal of Social Issues*, 46, 125–139.
- Rubovitz, R., & Maehr, M. (1973). Pygmalion black and white. *Journal of Personality and Social Psychology*, 19, 197–203.
- Ryan, C. (1995). Motivations and the perceiver's group membership: Consequences for stereotype accuracy. In Y. T. Lee, L. Jussim, & C. R. McCauley (Eds.), *Stereotype accuracy* (pp. 189–214). Washington, DC: American Psychological Association.
- Ryan, C. (1996). Accuracy of Black and White college students' in-group and out-group stereotypes. *Personality and Social Psychology Bulletin*, 22, 1114–1127.
- Ryan, C. S. (2002). Stereotype accuracy. *European Review of Social Psychology*, 13, 75–109.
- Ryan, C. S., & Bogart, L. M. (2001). Longitudinal changes in the accuracy of new group members' in-group and out-group stereotypes. *Journal of Experimental Social Psychology*, 37, 118–133.
- Ryan, C., Park, B., & Judd, C. M. (1996). Assessing stereotype accuracy: Implications for understanding the stereotyping process. In C. N. Macrae, C. Stangor, & M. Hewstone (Eds.), *Stereotypes and stereotyping* (pp. 121–150). New York: Guilford.
- Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2004). On interpreting stereotype threat as accounting for African-American–White differences on cognitive tests. *American Psychologist*, 59, 7–13.
- Sackett, P. R., & Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist*, 49, 929–954.
- Sagar, H. A., & Schofield, J. W. (1980). Racial and behavioral cues in black and white children's perceptions of ambiguously aggressive acts. *Journal of Personality and Social Psychology*, 39, 590–598.
- Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124, 262–274.

- Schneider, D. J. (2004). *The psychology of stereotyping*. New York: Guilford Press.
- Schneider, D. J., Hastorf, H., & Ellsworth, P. C. (1979). *Person perception* (2nd ed.). Reading, MA: Addison-Wesley.
- Schultz, P. W., & Oskamp, S. (2000). *Social psychology: An applied perspective*. Upper Saddle River, NJ: Prentice-Hall.
- Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys*. New York: Academic Press.
- Sedikides, C., & Skowronski, J. (1991). On the law of cognitive structure activation: Reply to commentaries. *Psychological Inquiry*, 2, 211–219.
- Shelton, N. (2000). A reconceptualization of how we study issues of racial prejudice. *Personality and Social Psychology Review*, 4, 374–390.
- Sidanius, J., & Pratto, F. (1999). *Social dominance: An intergroup theory of social hierarchy and oppression*. New York: Cambridge University Press.
- Siegel, J. J., & Bernstein, P. L. (1998). *Stocks for the long run* (2nd ed.). New York: McGraw-Hill.
- Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher's z transformation be used? *Journal of Applied Psychology*, 72, 146–148.
- Simon, B., & Pettigrew, T. F. (1990). Social identity and perceived group homogeneity: Evidence for the ingroup homogeneity effect. *European Journal of Social Psychology*, 20, 269–286.
- Skinner, B. F. (1990). Can psychology be a science of mind? *American Psychologist*, 45, 1206–1210.
- Skov, R. B., & Sherman, S. J. (1986). Information-gathering processes: Diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation. *Journal of Experimental Social Psychology*, 22, 93–121.
- Skrypnck, B. J., & Snyder, M. (1982). On the self-perpetuating nature of stereotypes about women and men. *Journal of Experimental Social Psychology*, 18, 277–291.
- Smedley, J., & Bayton, J. (1978). Evaluative race-class stereotypes by race and perceived class of subjects. *Journal of Personality and Social Psychology*, 30, 530–535.
- Smith, A., Jussim, L., & Eccles, J. (1999). Do self-fulfilling prophecies accumulate, dissipate, or remain stable over time? *Journal of Personality and Social Psychology*, 77, 548–565.
- Snow, R. E. (1969). Unfinished Pygmalion. *Contemporary Psychology*, 14, 197–200.
- Snow, R. E. (1995). Pygmalion and intelligence? *Current Directions in Psychological Science*, 4, 169–171.
- Snyder, M. (1984). When belief creates reality. *Advances in Experimental Social Psychology*, 18, 247–305.
- Snyder, M. (1992). Motivational foundations of behavioral confirmation. *Advances in Experimental Social Psychology*, 18, 247–305.
- Snyder, M., & Miene, P. (1994). On the functions of stereotypes and prejudice. In M. P. Zanna & J. M. Olson (Eds.), *The psychology of prejudice: The Ontario symposium* (pp. 33–54). Hillsdale, NJ: Erlbaum.
- Snyder, M., & Stukas, A. A., Jr. (1998). Interpersonal processes: The interplay of cognitive, motivational, and behavioral activities in social interaction. *Annual Review of Psychology*, 50, 273–303.
- Snyder, M., & Swann, W. B. (1978a). Behavioral confirmation in social interaction: From social perception to social reality. *Journal of Experimental Social Psychology*, 14, 148–162.
- Snyder, M., & Swann, W. B., Jr. (1978b). Hypothesis-testing processes in social interaction. *Journal of Personality and Social Psychology*, 36, 1202–1212.

- Snyder, M., Tanke, E. D., & Berscheid, E. (1977). Social perception and interpersonal behavior: On the self-fulfilling nature of social stereotypes. *Journal of Personality and Social Psychology*, 35, 656-666.
- Snyder, M., & Uranowitz, S. W. (1978). Reconstructing the past: Some cognitive consequences of person perception. *Journal of Personality and Social Psychology*, 36, 941-950.
- Sommers, S. R., & Ellsworth, P. C. (2000). Race in the courtroom: Perceptions of guilt and dispositional attributions. *Personality and Social Psychology Bulletin*, 26, 1367-1379.
- Sommers, S. R., & Norton, M. I. (2007). Race-based judgments, race-neutral justifications: Experimental examination of peremptory use and the *Batson* challenge procedure. *Law and Human Behavior*, 31, 261-273.
- Spitz, H. H. (1999). Beleaguered Pygmalion: A history of the controversy over claims that teacher expectancy raises intelligence. *Intelligence*, 27, 199-234.
- Stangor, C. (1995). Content and application inaccuracy in social stereotyping. In Y. T. Lee, L. Jussim, & C. R. McCauley (Eds.), *Stereotype accuracy* (pp. 275-292). Washington, DC: American Psychological Association.
- Stangor, C., & McMillan, D. (1992). Memory for expectancy-congruent and expectancy-incongruent information: A review of the social and social developmental literatures. *Psychological Bulletin*, 111, 42-61.
- Steele, C. M. (1992, April). Race and the schooling of black Americans. *Atlantic Monthly*, 68-78.
- Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, 52, 613-629.
- Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual performance of African Americans. *Journal of Personality and Social Psychology*, 69, 797-811.
- Stephan, W. G., & Stephan, C. W. (1985). Intergroup anxiety. *Journal of Social Issues*, 41, 157-175.
- Stevens, S., Cohen, F., & Jussim, L. (2008). The role of racism and sexism in the 2008 presidential election. Unpublished manuscript.
- Stinson, L., & Ickes, W. (1992). Empathic accuracy in the interactions of male friends versus male strangers. *Journal of Personality and Social Psychology*, 62, 787-797.
- Sutherland, A., & Goldschmid, M. L. (1974). Negative teacher expectation and IQ change in children with superior intellectual potential. *Child Development*, 45, 852-856.
- Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. New York: Doubleday.
- Swann, W. B. (1984). Quest for accuracy in person perception: A matter of pragmatics. *Psychological Review*, 91, 457-477.
- Swann, W. B., Jr. (1987). Identity negotiation: Where two roads meet. *Journal of Personality and Social Psychology*, 53, 1038-1051.
- Swann, W. B., Jr., & Ely, R. J. (1984). A battle of wills: Self-verification versus behavioral confirmation. *Journal of Personality and Social Psychology*, 46, 1287-1302.
- Swann, W. B., Jr., & Giuliano, T. (1987). Confirmatory search strategies in social interaction: How, when, why, and with what consequences. *Journal of Social and Clinical Psychology*, 5, 511-524.
- Swann, W. B., Jr., Giuliano, T., & Wegner, D. M. (1982). Where leading questions can lead: The power of conjecture in social interaction. *Journal of Personality and Social Psychology*, 42, 1025-1035.

- Swann, W. B., Griffin, J. J., Predmore, S., & Gaines, B. (1987). The cognitive-affective crossfire: When self-consistency confronts self-enhancement. *Journal of Personality and Social Psychology*, 52, 881–889.
- Swann, W. B., Jr., Milton, L. P., & Polzer, J. T. (2000). Should we create a niche or fall in line? Identity negotiation and small group effectiveness. *Journal of Personality and Social Psychology*, 79, 238–250.
- Swann, W. B., Jr., & Read, S. J. (1981a). Acquiring self-knowledge: The search for feedback that fits. *Journal of Experimental Social Psychology*, 17, 1119–1128.
- Swann, W. B., Jr., & Read, S. J. (1981b). Self-verification processes: How we sustain our self-conceptions. *Journal of Experimental Social Psychology*, 17, 351–372.
- Swim, J. K. (1994). Perceived versus meta-analytic effect sizes: An assessment of the accuracy of gender stereotypes. *Journal of Personality and Social Psychology*, 66, 21–36.
- Swim, J., Borgida, E., Maruyama, G., & Myers, D. G. (1989). Joan McKay vs. John McKay: Do gender stereotypes bias evaluations? *Psychological Bulletin*, 105, 409–429.
- Taft, R. (1955). The ability to judge people. *Psychological Bulletin*, 52, 1–23.
- Tajfel, H. (1981). *Social identity and intergroup relations*. London: Cambridge University Press.
- Taylor, M. C. (1979). Race, sex, and the expression of self-fulfilling prophecies in a laboratory teaching situation. *Journal of Personality and Social Psychology*, 37, 897–912.
- Taylor, M. C. (1992). Expectancies and the perpetuation of racial inequality. In P. D. Blanck (Ed.), *Interpersonal expectancies*. New York: Cambridge University Press.
- Taylor, S. E., Fiske, S. T., Etcoff, N. L., & Ruderman, A. (1978). Categorical and contextual bases of person memory and stereotyping. *Journal of Personality and Social Psychology*, 36, 426–432.
- Terracciano, A., Abdel-Khalek, A. M., Adam, N., Ladmovova, L., Ahn, C.-k, Ahn, H.-n, et al. (2005). National character does not reflect mean personality trait levels in 49 cultures. *Science*, 310, 96–100.
- Tetlock, P. E. (2002). Social-functionalist frameworks for judgment and choice: The intuitive politician, theologian, and prosecutor. *Psychological Review*, 109, 451–472.
- Thorndike, R. L. (1968). Review of Pygmalion in the classroom. *American Educational Research Journal*, 5, 708–711.
- Trope, Y., & Bassok, M. (1982). Confirmatory and diagnosing strategies in social information gathering. *Journal of Personality and Social Psychology*, 43, 22–34.
- Trope, Y., & Bassok, M. (1983). Information-gathering strategies in hypothesis-testing. *Journal of Experimental Social Psychology*, 19, 560–576.
- Trope, Y., Bassok, M., & Alon, E. (1984). The questions lay interviewers ask. *Journal of Personality*, 52, 90–106.
- Trouilloud, D., Sarrazin, P., Martinek, T., & Guillet, E. (2002). The influence of teacher expectations on students' achievement in physical education classes: Pygmalion revisited. *European Journal of Social Psychology*, 32, 1–17.
- U.S. Census (2010a). Education. Retrieved 17 November, 2011 from <http://www.census.gov/prod/2009pubs/10statab/educ.pdf>.
- U.S. Census (2010b). Income, expenditures, poverty, and wealth. Retrieved 17 November, 2011 from <http://www.census.gov/prod/2009pubs/10statab/income.pdf>.
- U.S. News & World Report (September 22, 2008). Poll: Barack Obama could lose six percentage points on Election Day for Being Black. Retrieved on 11/26/11 from: <http://www.usnews.com>.

- com/news/campaign-2008/articles/2008/09/22/poll-barack-obama-could-lose-six-percentage-points-on-election-day-for-being-black
- von Baeyer, C. L., Sherkr, D. L., & Zanna, M. P. (1981). Impression management in the job interview: When the female applicant meets the male (chauvinist) interviewer. *Personality and Social Psychology Bulletin*, 7, 45–51.
- von Hippel, W. (2004). Implicit prejudice: Pentimento or inquisition? *Psychological Inquiry*, 15, 302–305.
- Wall Street Journal. (1994). Mainstream Science on Intelligence. Reprinted in Gottfredson, L.S. (1997), Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence*, 24, 13–23.
- Wegner, D. M., & Vallacher, R. R. (1977). *Implicit psychology: An introduction to social cognition*. New York: Oxford University Press.
- Weinstein, R. S., Gregory, A., & Strambler, M. J. (2004). Intractable self-fulfilling prophecies: Fifty years after Brown v. Board of Education. *American Psychologist*, 59, 511–520.
- Weinstein, R. S., & McKown, C. (1998). Expectancy effects in “context”: Listening to the voices of students and teachers. In J. Brophy (Ed.), *Advances in research on teaching*, Vol. 7 (pp. 215–242). Greenwich, CT: JAI Press.
- West, C., & Anderson, T. (1976). The question of preponderant causation in teacher expectancy research. *Review of Educational Research*, 46, 613–630.
- Wigfield, A., Eccles, J. S., MacIver, D., Reuman, D., & Midgley, C. (1991). Transition at early adolescence: Changes in children’s domain-specific self-perceptions and general self-esteem across the transition to junior high school. *Developmental Psychology*, 27, 552–565.
- Wijmenga, R. T. (1990). The performance of published Dutch stock recommendations. *Journal of Banking and Finance*, 14, 559–581.
- Wilder, D. A. (1986). Social categorization: Implications for creation and reduction of intergroup bias. *Advances in Experimental Social Psychology*, 19, 291–355.
- Williams, T. (1976). Teacher prophecies and the inheritance of inequality. *Sociology of Education*, 49, 223–236.
- Wineburg, S. S. (1987). The self-fulfillment of the self-fulfilling prophecy: A critical appraisal. *Educational Researcher*, 16, 28–40.
- Winston Churchill Leadership. Winston Churchill on others... Retrieved November 24, 2008, from <http://www.winston-churchill-leadership.com/churchill-quote-others.html>
- Wolsko, C., Park, B., Judd, C. M., & Wittenbrink, B. (2000). Framing interethnic ideology: Effects of multicultural and color-blind perspectives on judgments of groups and individuals. *Journal of Personality and Social Psychology*, 78, 635–654.
- Word, C. O., Zanna, M. P., & Cooper, J. (1974). The nonverbal mediation of self-fulfilling prophecies in interracial interaction. *Journal of Experimental Social Psychology*, 10, 109–120.
- Ysseldyke, J. E., Algozzine, B., Shinn, M., & McGue, M. (1982). Similarities and differences between low achievers and students classified as learning disabled. *Journal of Special Education*, 49, 223–236.
- Zanna, M. P., & Pack, S. J. (1975). On the self-fulfilling nature of apparent sex differences in behavior. *Journal of Experimental Social Psychology*, 11, 583–591.
- Zuckerman, M., Knee, C. R., Hodgins, H. S., & Miyake, K. (1995). Hypothesis confirmation: The joint effect of positive test strategies and acquiescence response set. *Journal of Personality and Social Psychology*, 68, 52–60.

This page intentionally left blank

Index

Note: Page numbers followed by “*f*”, “*n*”, or “*t*” refer to figures, notes, or tables, respectively.

- absolute discrepancies, 356
- absolutist stereotypes, 312
- academic achievement, sex stereotyping
 - and, 343–44
- accumulation
 - accuracy and, 252
 - attractiveness and, 261–62
 - concurrent, 253–58, 256*f*, 263
 - limitations to, 252–53
 - logic of, 251–52
 - MBAs and, 261–63
 - over time, 258–59
 - Pygmalion study and, 259
 - regression to mean and, 253
 - of self-fulfilling prophecies, 248–65
 - self-verification and, 252
 - of small bias, 375
 - synergistic, 257–58
- accuracy, 6, 8, 11, 22, 168, 246*n*6,
 - 417–19, 422. *See also* inaccuracy;
 - stereotype accuracy
- assessment of, 161
- of beliefs, 367
- bias compared with, 423
- cognitive process research and, 148–50, 157–59
- of consensual stereotypes, 339, 340, 342, 395–96
- constructivist perspectives on, 173–74
- construct validity and, 206–9
- content, 418
- correlational approaches, 205–6
- correspondence and, 172–73, 181
- criterion problem, 165–67
- Cronbach on, 147–48
- data on, 370–88
- differential, 196–97, 215*n*6
- double standards, 166–67
- dyadic, 199–200, 321*n*1
- elevation, 195, 198, 200
- empathic, 417–18
- establishing, 202–3, 208
- of explanations, 159–60
- Fiske on, 154–55
- full accuracy design 200–201
- functional perspectives, 173
- generalized, 199
- history of, 145–47
- improper linear models, 211–13
- inaccuracy and, 168*n*3, 295
- independent of influence, 172–73
- individuating information and, 366–70

- accuracy, (*contd.*)
- inequality and, 152
 - Jones, E. E., on, 149–50
 - Judd and Park on, 200–201
 - legality of, 155–56
 - lower bounds, 182
 - in New Look, 22
 - noncomponential approaches, 204–5, 392
 - objections to, 145–69
 - overall levels of, 178
 - overestimating, 367
 - perceiver, 198
 - personal stereotype, 329–30
 - political objections to, 152–55
 - presidential elections and, 191*n*1
 - problems with, 158–59
 - process compared to, 201–2
 - Realistic Accuracy Model, 210–13
 - research, 147–52, 154–55
 - scientific validity of, 163–64
 - self-fulfilling prophecy accumulation and, 252
 - self-fulfilling prophecy compared with, 160–65
 - of sex stereotypes, 335–48
 - social beliefs and, 169*n*8
 - social problem alleviation and, 156–57
 - social psychology on, 149
 - of social stereotypes, 393*t*
 - theoretical objections to, 157
 - thick slices and, 419
 - from thin slices, 418–19
 - truth and, 175–76
 - types of, 397
 - underestimating, 181
 - valence, 418
 - at zero acquaintance, 418–19
- accuracy components, 194–216
- assessing, 201–4
 - Cronbach's, 195–97, 215*n*2
 - models, 203*t*, 204
- accuracy criteria, 170–93
- behavior, 183–84
 - construct validity and, 175
 - establishing, 175–91
 - hybrid, 190–91
 - imperfection of, 191
 - objective, 18, 176–82
 - probabilistic realism, 171–73
 - theoretical perspectives on, 171–75
 - truth and, 175–76
- acquiescence, 119–20
- ad hominem attack, 155
- adjusted means, 419*n*1
- African Americans, 53, 99*n*3, 178, 238, 371
- Duncan study, 126–29
 - hostility and, 410–11
 - self-fulfilling prophecy and, 26–27
 - stereotypes, 324–27, 411
 - stereotype threat and, 408–9
 - student perceptions, 328–31, 419*n*1
- agreement
- behavior and, 183–84
 - bias in, 187
 - with expert models, 184–86
 - with experts, 184, 185
 - with independent judges, 185
 - limitations to, 185–86, 189
 - with nonindependent judges, 185
 - with perceivers, 184–87
 - self-fulfilling prophecies in, 187
 - with target, 187–89
 - with target self-perceptions, 188–89
 - with target self-reports, 187–88
- Al Qaeda, 365
- Alaska, 362, 364, 366, 426
- Albright, L., 94
- social class stereotype study, 129–31
- alcohol use, 257–58
- Allport, F., 19–22, 24
- on New Look, 22
- Allport, Gordon, on stereotyping, 18–19, 283, 295, 313, 423
- on stereotypes, 285–86, 293, 303–4
- Alon, E., information-seeking study, 114
- Alvidrez, J., on self-fulfilling prophecies, 243–44
- Alwin, D., 245*n*3
- Amabile, T. M., 78*n*1
- Amazon.com, 89
- Ambady, N., 419
- ambiguous information, 363, 379
- Ambivalent Sexism Inventory
- (Glick and Fiske), 341
- American Psychological Association
- (APA), 358–59
- American Psychological Society, 305*n*2
- anagrams test, 56
- analysis of covariance (ANCOVA), 419*n*1

- analysis of variance (ANOVA)
 Cronbach's components as, 197
 Kenny's components as, 200
 ANCOVA. *See* analysis of covariance
 Anderson, S. M., 96, 260
 on teacher expectations, 223
 animal subjects
 experimenter effects on, 33
 mice, 284
 ANOVA. *See* analysis of variance
 anthropology, 151
 antidiscrimination lawsuits, 275, 306*n*5
 APA. *See* American Psychological Association
 appraisal effects. *See* self-fulfilling prophecies
 Armenians, 16–18
 Aronson, E., 409, 419*n*1
 Asch, S., 24–25
 Ashmore, Richard, 127–28, 304
 on stereotypes, 302
 Ashton, M. C., 347, 399
 racial stereotype study, 325*t*–326*t*, 331–33
 Asian Americans, 358
 assertiveness, 372
 attention deficit/hyperactivity disorder, 135
 attitudes, 410
 attractiveness. *See* physical attractiveness
 stereotype
 attribute effect, 200
 attributional bias, 50*t*, 51*t*
 attributions, 410
 authoritarianism, 169*n*6
 automaticity, 410, 411

 Babad, E., 241
 ballet dancer stereotypes, 352–54
 bank insolvency, self-fulfilling prophecies
 and, 90–91
 Bargh, J. A., 54
 racial stereotype study, 410–12
 on self-fulfilling prophecy, 99*n*3
 Baron, R. M., 94
 social class stereotype study, 129–31
 baseball, 157, 194, 215*n*1, 215*n*3,
 215*n*4, 215*n*5
 base-rates, 370
 basic construct validity, 206
 Bassok, M.
 on confirmation, 117
 on diagnosis, 117
 on hypothesis-testing, 117
 information-seeking study, 114
 battle of wills, 107–8, 120
 behavior
 accuracy criteria, 183–84
 agreement and, 183–84
 defined, 183
 hostile, 60–61, 407, 411
 nonverbal, 339–40
Behavioral and Brain Sciences, 40
 behavioral effects of expectancy, 50*t*
 behaviorism, 148, 151
 beliefs, 9
 accuracy of, 367
 hostility and, 60–61
 moral, 9
 political, 9, 152–56, 296–97, 346–48, 424
 prescriptive, 9
 similarity model, 371
 social, 11–12, 169*n*8, 172
The Bell Curve (Herrnstein & Murray), 37
 Bellezza, F. S., 139
 Bem, S. L., 96
 benevolence, 245*n*4
 benevolent sexism subscale, 341
 Berscheid, E., 58, 294
 critical evaluation of, 95–97
 BESD. *See* binomial effect size display
 Beyer, S., 402
 sex stereotype study, 337*t*, 343–46
 bias, 4–7, 10, 50*t*, 168, 214, 413–17, 427
 accuracy compared with, 423
 in agreement, 187
 assessment of, 168*n*4
 attributional, 50*t*, 51*t*
 bias against, 377
 bias in favor of, 400–401
 cognitive psychology and, 12*t*
 contraction, 352
 data on, 370–88
 dominated by, 29*n*2
 empathic accuracy and, 418
 evaluative, 50*t*, 51*t*
 expectancy-confirming, 123–25
 expectancy-induced, 76–77
 expectations, information gathering, 73–74
 expectations, person perception, 65–70,
 72–73
 finding, 413

- bias, (*Contd.*)
- gender, 137
 - in Hastorf and Cantril, 23–24
 - hypothesis, 118–21
 - inaccuracy and, 306*n*6
 - information-seeking, 51*t*
 - judgmental, 50*t*, 51*t*, 85*t*
 - jury selection, 413–15
 - limitations, 416
 - in math classes, 227
 - memory, 50*t*, 51*t*, 67–68, 85*t*
 - in New Look, 22
 - power of, 414–17
 - in questions, 114
 - reasonableness and, 416
 - replication of, 416
 - response, 195
 - sex stereotype, 387
 - small, accumulation, 375
 - social psychology and, 12*t*, 169*n*5
 - stereotype and, 18–19, 290–91
 - stereotypes, person perception, 65–70
 - in Terracciano study, 400–401
 - thick slices and, 419
 - in urban schools, 226
- Biden, Joseph, 432*n*2
- Big Five personality traits, 388*n*2
- binomial effect size display (BESD), 228, 320, 392–93
- bipolar disorder, 135
- Bodenhausen, 373
- Bogart, 350*t*
- sorority study, 355–56
- boomerang effect, 96
- Borgida, 294
- on individuating information, 372–74
- Bower, G. H., 139
- Braly, K., 18, 28, 66, 294
- on stereotyping, 15–16, 313, 315, 317–18
- Brattesani, K. A., 223–24
- Brekke, K. A., 294
- on individuating information, 372–74
- Brewer, M. B., 304
- Brigham, J. C., 301, 304
- British, 331–33
- Briton, N. J., 339–40
- sex stereotype study, 336*t*
- Brodt, S., 305*n*2
- on person perception, 384
- Brophy, J., 228, 244
- Brunswick, E., lens model, 145–46, 209–10, 209*f*, 212–13
- bull's eye, 318
- business major stereotypes, 353–54
- Campbell, D. T., 304, 371, 372
- person perception, 382
- Canadian-born blacks, 331–33
- Cantril, H., 21–22, 131, 313, 423
- bias in, 23–24
- Caribbean-born blacks, 331–33
- Carnegie, Dale, 64
- Carter, 399
- sex stereotype study of, 336*t*, 340–42
- caste system, 92
- causal language, 246*n*6
- Cejka, M. A., 337*t*, 348–52, 349*t*
- cell phones, 88
- Census, U.S., 178, 315, 327
- central traits, 25
- Chapman, Gretchen, 216*n*8, 225*t*
- charter schools, 305*n*3
- Chein, L., 20
- Chen, M., 54
- racial stereotype study, 410–12
 - on self-fulfilling prophecy, 99*n*3
- Chinese, 331–33, 371, 376
- Churchill, Winston, 377
- Cisco, 89
- civil rights law, 155, 305*n*2, 368, 402
- civil rights movement, 34
- Clabaugh, A., 350*t*, 397
- ballet dancer stereotype study, 352–54
- Claire, T., on stereotypes, 253–54, 375–76
- Clarke, L. F., 139
- Clarke, R. B., 371
- person perception, 382
- classroom
- dissipation in, 259–60
 - expectancy-confirming biases in, 224–25
 - self-fulfilling prophecies, 222–24
 - sex stereotypes in, 56–57
 - social class in, 259–60
- coaching, soccer, 123–24
- cognitive ability tests, 179–81
- cognitive dissonance, 410
- cognitive effects of expectancy, 50*t*
- cognitive judgment, 187

- cognitive misers, 4, 383, 417
- cognitive process research
 accuracy and, 148–50, 157–59
 in social perception, 150
- cognitive psychology, 11–12, 151–52
 bias and, 12*t*
- Cohen, C. E., 67–68, 294, 320
 memory study, 138–41
 on person perception, 382–83
- college admissions, 411–12
- Collins, Barry, 127–28
- color-blind perspective, 333–35, 341, 399
- competence, student, 234
- competition, self-fulfilling, 59
- componential models, 418
 Judd and Park, 392
 of self-judgment, 203*t*
 self-perceptions, 204
 self-reports, 204
- concurrent accumulation, 253–58, 256*f*, 263
 empirical evidence of, 257
 overestimating, 256
 underestimating, 256
- confirmation
 Devine on, 118
 diagnosis compared with, 117
 in hypothesis-testing, 117
 in information seeking, 115–18
 Skov and Sherman on, 118
 Trobe and Bassok on, 117
- confirmatory questions, 115
- conscientiousness, 400
- consensual stereotypes, 317–18, 321
 accuracy of, 324, 325–26*t*, 327, 328–30,
 331–33, 334, 335, 336–38*t*, 339, 340,
 342, 343, 344–45, 346–47, 349–51*t*,
 352–53, 354, 395–96
 correspondence, 327, 345, 356
 discrepancies, 326*t*, 327, 328–29, 331,
 334, 335, 339, 342, 343, 344, 346–47,
 353, 354, 356
 with objective criteria, 335–39
 personal stereotypes compared to, 395–96
 racial, 327
 in Ryan study, 329
 sex, 335–39
- conservatives, 151, 152
- constrained perceivers, 113–14
- constraining questions, 113–14
- constructivist perspectives of accuracy, 173–74
- constructs, defining, 291
- construct validity
 accuracy and, 206–9
 accuracy criteria and, 175
 basic, 206
 of extroversion, 207
 of intelligence, 207
- contact hypothesis, 398
- Contemporary Psychology*, 38
- content, 417
 accuracy, 418
 patterns for, 418
- contraction bias, 352
- contrary indicators, 88
- Cooper, H., 53
- Copus, D., 299
- correlational approach, 205–6
- correlational criterion, 189
- correlations, 141*n*3
- correspondence, 172–73
 consensual stereotypes, 327, 345, 356
 with differences, 317, 320–21
 lack of, 181
 personal stereotypes, 333, 345, 347, 398
- counterstereotypic attributes, 330
- Cozzarelli, C., 105–7
- criterion problem, 165–67
- Cronbach, L. J., 168*n*2, 200–203, 205, 213,
 215*n*6, 418
 on accuracy, 147–48
 accuracy components, 195–97, 215*n*2
 Jones, E. E., on, 149
- cross-cultural psychology, 151
- crowds, 395–96
- cultural myths, false, 396
- cultural psychology, 288–89
- culture, 6
- curvilinear relationships, 43
- Darley, J. M., 67, 94, 142*n*5, 423
 social class stereotype study, 129–31
- data
 on accuracy, 370–88
 balanced social sciences and, 426–28
 on bias, 370–88
 on error, 370–88
 paying attention to, 367
 on reasonableness, 370–88

- Dawes, R. M., 216*n*9
 improper linear models, 211–13
- Deaux, K., 67
- decision making, 152, 187
 models, 193*n*6
- definitive individuating information, 362–63
- Del Boca, F. K., on stereotypes, 302
- Democrats, 354, 431 (Democratic candidates)
- denial of difference ideology, 397
- depression, 135
- Devine, P., 304
 on confirmation, 118
 on diagnosis, 118
- diagnosis
 confirmation compared with, 117
 in hypothesis-testing, 117
 in information seeking, 115–18
 Trope and Bassok on, 117
- Diagnostic and Statistical Manual of the Mental Disorders* (DSM-IV), 133
- diagnostic information, 115
- Diekmann, A. B., 303, 402
 sex stereotype study, 337*t*, 346–48
- differential accuracy, 196–97, 215*n*6
- differential elevation, 196, 215*n*6
 accuracy, 196
- disconfirmation, 115–17
 falsification compared with, 116
- discrepancy
 absolute, 356
 consensual stereotypes, 326*t*, 327, 331, 334, 335, 339, 344, 347
 from perfection, 316–17
 personal stereotype, 331–32, 347
 scores, 353–56
 stereotype accuracy, 318–20
 types, 319–20
- discrimination
 antidiscrimination lawsuits, 275, 306*n*5
 group differences and, 369
 stereotypes and, 299, 391
- disparate impact, 156
- dissipation
 in classroom, 259–60
 in elementary school, 260
 incomplete, 260–61
 in Pygmalion, 259
- diversity, 369
- double standards
 in accuracy and self-fulfilling prophecies, 166–67
 in cognitive ability tests, 179–81
- Dow Jones Industrial, 90, 192*n*4
- Downey, G., 412
- duck test, 175–76
- Duncan, B. L., 66, 126–28
- dyadic accuracy, 199–200, 321*n*1
- dyadic interactions, 255
- Eagly, A. H., 303, 346–48, 402
 sex stereotype study, 337*t*, 346*t*, 348–52
- earned reputation theory. *See* exaggeration hypothesis
- East Indians, 331–33
- Eccles, J. S., 237–38, 240
 on teacher expectations, 227
- Eden, Dov, 99*n*1
- educational psychology, 79*n*4
- efficient markets theory, 186
- Effrein, E. A., 75–76
- egalitarian denial hypothesis, 398–99, 415
- egalitarian movements, 303, 398
- Ekman, P., 178
- elementary school, dissipation of
 self-fulfilling prophecies through, 260
- elevation, 418
 accuracy, 195, 198, 200
 differential, 196, 215*n*6
 scores, 198
- Ely, R. J.
 predictions from, 103*t*
 on self-verification, 102–5, 109, 110, 120–21
- empathic accuracy
 bias and, 418
 content in, 417–18
 error and, 418
 valence in, 417–18
- Emswiller, T., 67
- engineering major stereotypes, 353–54
- erroneous expectation, 411
- error, 4–5, 10, 11
 data on, 370–88
 empathic accuracy and, 418
 inaccuracy and, 306*n*6
 overdemonstration of, 423
 reign of, 61
 in stereotypes, 290–91
 thick slices and, 419

- Esses, V. M., 347, 399
 racial stereotype study, 325*t*–326*t*, 331–33
- Etcoff, N. L., 294
- evaluative bias, 50*t*, 51*t*
- Evans, M., 71–72
- evil, stereotypes and, 279–80
- evolutionary biology, 427
- exaggeration hypothesis, 314, 392, 398
- expectancies (see also expectations), 8, 393, 410
 behavioral effects of, 50*t*
 classes of effects, 50*t*
 cognitive effects of, 50*t*
 confirming responses, 120
 critical analysis of, 83–84
 disconfirming, 104
 effect size, 43*f*; 85*t*–86*t*
 effects of, 49–51
 induced biases, 76–77
 induction, 43*f*
 injustice and, 153
 research, 421
- expectancy-confirming bias, 123–25
 in classroom, 224–25
 racial stereotype studies, 126–29
 on sex stereotypes, 131–32
 social class stereotypes, 129–31
- expectancy-inconsistent information, 139–40
- expectations, 30. *See also* teacher expectations
 bias and, 122–42
 bias information gathering, 73–74
 bias person perception, 65–70, 72–73
 clinging to, 106
 erroneous, 411
 false, 254
 inaccurate, 77–78
 information seeking and, 112–21
 interpersonal, 63*n*2
 judgment and, 122–42
 memory and, 122–42
 perceivers, 162, 255–56
 perception and, 122–42
 positive, 37
 self-perception and, 408
 stereotype-based, 52, 125–41
 stereotypes as biased, 66
- experimental psychology, 33
- experimental studies, teacher expectations, 222
- experimenter effects
 on animal subjects, 33
 on human subjects, 32–33
 Rosenthal's work on, 31–33
- expert models, agreement with, 184–86
- experts, agreement with, 184, 185
- explanations, 159–60
- Extraordinary Popular Delusions and the Madness of Crowds* (MacKay), 192
- extroversion, 103, 114, 119–20, 384
 as construct, 207
 testing for, 116
- fads, 395, 410
- fake males study, 54–55
- Falender, V. J., 75–76
- false cultural myths, 396
- false expectations, 254
- falsifiability, 163
- falsification, 115–17
 disconfirmation compared with, 116
- Fazio, R. H., 75–76
- Feingold, A., sex stereotype meta-analysis, 137
- Finn, J., on bias in urban schools, 226
- Fisher's *r*-to-*z* transformation, 359*n*1
- Fiske, S. T., 169*n*7, 294, 304, 305*n*2, 341, 373, 377, 427, 428
 on accuracy, 154–55
 on individuating information, 372–73
 on stereotypes, 253–54, 297, 298, 375–76
- Flanagan's Test of General Ability (TOGA), 34
- Fode, K. L., 33
- Friesen, W. V., 178
- Frieze, I. H., 263, 265*n*3, 265*n*4
 on MBAs and accumulation, 261–62
- Fulero, S., friendly/intelligent memory study of, 71–72
- full accuracy design, 200–201
- functional perspectives of accuracy, 173
- Funder, D. C., 175
 Realistic Accuracy Model, 210–13
- future predictions, 363–64
- gang members, 365
- Gary, Mel, 127–28
- gays, 139
- gender, 384, 385. *See also* sex stereotypes
 bias, 137
- generalizations, 300
- generalized accuracy, 199
- General Social Survey, 346

- Ghazarian, S. R., on self-fulfilling prophecies, 241
- Gilovich, T., 11
- Giuliano, T., 114–15
- glass half-full parable, 3–4, 421–32
- Glick, P., 341
- Goldberg, P., 67
sex stereotypes study, 131–32
- Goldschmid, M. L., on self-fulfilling prophecies, 242–43
- Gore, Al, 177
- Gosling, S. D., on person perception, 384–85, 387–88
- Great Depression, 26
- Greenspan, Alan, 8
- GREs, 45, 46*n*5, 180, 212, 216*n*8, 255, 374
- Gross, P. H., 67, 94, 142*n*5, 423
social class stereotype study, 129–31
- group differences, 284–87
in cultural psychology, 288–89
denial of, 398
discrimination and, 369
in person perception, 386
social sciences on, 391
sources of, 391
- group effect
perceiver, 200
target, 200
- group-serving attributions, 393
- Guillet, E., teacher expectation study, 224
- Guyll, M., on synergistic accumulation, 257–58
- Ha, Y., 119
- Hall, J. A., 339–42, 399
sex stereotype study, 336*t*
- Hamilton, D. L., 304
on stereotypes, 297
- Handbook of Social Psychology*, 148
- Hastorf, A. H., 21–22, 132, 313, 423
bias and, 23–24
- Hauser, R. M., 245*n*3
- Heine, S. J., 399
national personality stereotypes study, 399–400
- Hepburn, C., 294
on individuating information, 372–74
- Herrnstein, R. J., 37, 156
- heuristics, 393
- high-certainty perceivers, 104–5
- high-certainty targets, 103–4
- high wattage view, 4–5
- Hinnant, J. B., 225*t*, 260, 261
on self-fulfilling prophecies, 241
- Hodgins, H. S., on bias hypothesis, 119–21
- Hofer, Myron A., 37
- home field advantage, 87
- Homo sapiens*, 5
- hostile behavior, 60–61, 407
African American stereotype of, 411
- hostile sexism subscale, 341
- hostility, self-fulfilling beliefs about, 60–61
- How Schools Shortchange Girls*, 345
- How We Know What Isn't So* (Gilovich), 11
- Huff, Darrell, 38
- Human Inference: Strategies and Shortcomings of Social Judgment* (Nisbett & Ross), 11
- human subjects, 32–33
- Hunter, J. E., 181
- Huxley, Aldous, 323
- hybrid criteria, 190–91
- hypothesis-testing, 74, 79*n*5
bias, 118–21
confirmation in, 117
diagnosis in, 117
Trope and Bassok study on, 117
- Ickes, W., 417, 418
- ideology, 9
- IDF. *See* Israeli Defense Forces
- illusory correlations, 393
- imperfect criteria, 181
- implicit, 410
- implicit personality theory, 24–25, 146
- impressions
perceivers, 165
predictions and, 164–65
- improper linear models, 211–13
- inaccuracy, 6. *See also* accuracy
accuracy and, 168*n*3, 295
bias and, 306*n*6
conditions of, 401–2
discarding, 297–99
error and, 306*n*6
interpreting, 202
neutral definitions of, 297
in New Look, 22
Ryan, Park, and Judd on, 202

- stereotypes and, 16, 269–306, 391–92, 397–402
- inaccurate expectations, 77–78
- independent judges, agreement with, 185
- independent judgment, 395
- independent of influence, 172–73
- individual differences, 340–42
- individuals, judgment of, 360–70
 - individuating information in, 370–88
 - stereotypes accuracy in, 361
- individuating information, 362, 388*n*1, 428
 - abundant, 378–79
 - accuracy and, 366–70
 - ambiguous information in, 363, 379
 - definitive, 362–63
 - distortion of, 367
 - Fiske and Neuberg on, 372–73
 - future predictions, 363–64
 - in individual judgment, 370–88
 - inferences in, 363
 - lack of, 366–70, 379–80
 - Locksley, Borgida, Brekke, and Hepburn on, 372–74
 - meta-analyses on, 377
 - Obama and, 430
 - observations in, 363
 - past evaluation, 363–64
 - small amounts of, 363
 - useful, 363–65
- inequalities, 52
 - accuracy and, 152
 - self-fulfilling prophecies and, 421
 - stereotypes and, 280
- inferences, 363
- influence, independent of, 172
- information gathering
 - bias, 51*t*
 - confirmation in, 115–18
 - diagnosis in, 115–18
 - expectations and, 112–21
 - expectations bias, 73–74
 - self-fulfilling prophecies and, 75, 119–21
 - Snyder and Swan study on, 113–14
 - Zuckerman et al. study on, 119–21
- social, 118–21
- Swann and Giuliano study on, 114–15
- Trope, Bassok, and Alon study on, 114
- unbiased, 73–74
- in-group evaluations, 27–28
- injustice. *See* social injustice
- insolvency, bank, 90–91
- intellectual imperialism, 148–50
- intelligence, 5, 42
 - as construct, 207
 - of students, 234
- Internet, 88–89
- “Interpersonal Expectancy Effects: The First 345 Studies” (Rosenthal & Rubin), 52
- interpersonal perceptions, reality and, 6–7
- intrapsychic phenomenon, 148
- introversion, 103, 114
- introversion/extroversion study, 71
- IQ spurts, 226
- IQ tests, 46*n*5, 155, 175, 179, 180, 206, 242, 332. *See also specific tests*
 - Pygmalion study and, 33–39, 41–43, 45
 - Wineburg on, 42–43
- Irish, 376
- irrational exuberance, 8
- irrationality
 - in social psychology, 149–50
 - stereotypes and, 16, 375–77
- Israeli Defense Forces (IDF), 99*n*1
- Jackson, Jesse, 432*n*2
- Jacobson, L., 7, 45, 46, 51, 62, 190, 254
 - critical evaluation of, 92
 - on IQ spurts, 226
 - Pygmalion study of, 33–37, 39–40, 65, 220–21, 259, 424
- Japanese, stereotypes of, 139
- Jews, 272–73, 331–33, 376–77
- Jim Crow laws, 401
- job resumes, 367
- Jones, E. E., 148, 168*n*2, 304
 - on accuracy, 149–50
 - on Cronbach, 149
- Journal of Personality and Social Psychology*, 383
- Judd, C. M., 204, 328, 333
 - componential methodology, 392
 - full accuracy design, 200–201
 - on inaccuracy, 202, 317
 - political stereotypes study, 354–55
 - on relationship between stereotypes and prejudice, 298
 - sex stereotype study, 349*t*

- judgment
 cognitive, 187
 expectations and, 122–42
 of groups, 395
 independent, 395
 of individuals, 360–70
 of perceivers, 146
 rational, 416
 social, 203_t
 social reality and, 14
 stereotypes and, 150, 370–88
 unbiased, 416
- judgmental bias, 50_t, 51_t
 expectancy effect sizes and, 85_t
- jury selection, 413–15
- Jussim, L., 237–38, 240, 245_{n3}, 246_{n7}, 428
 on teacher expectations, 224, 227, 386
- Kahneman, Daniel, 11, 366, 370, 413
- Katz, D., 18, 28, 66, 294, 304
 on stereotyping, 15–16, 313, 315, 317–18
- Kelley, H. H., 59
 fundamental assumptions of, 146–47
 personal constructs theory, 146
 on social perception, 25–26
- Kelly, G. A., 176
 on people as naïve scientists, 146–47
- Kenny, D. A., 204, 213, 418
 social relations model, 197–200
- kernel of truth hypothesis. *See* exaggeration hypothesis
- Klayman, J., 119
- Knee, C. R., on bias hypothesis, 119–21
- known-groups validity, 287
- Ko, S. J., 384–85, 387–88
- Krueger, J., 294
- Kruglanski, A., 191
- Kuklinski, M. R., 225_t
- Kulesa, P., 346–48, 402
- Kulik, J. A., 71
- kumbaya hypothesis, 398–99, 415
- Kunda, Z., stereotype meta-analysis, 137, 378, 380
- labels, memory and, 138
- labor unions, 26–27
- LaPiere, R. T., 28, 28_{n1}, 66, 202, 304, 313, 372
 on stereotyping, 16–18
- The Last National Bank* (Merton), 26
- leadership, 396
- lens model, 145–46, 209–10, 212–13
 correlations, 209_f
- lesbians, 139
- Levine, R. A., 20
- liberals, 151, 152
- life transitions, 408
- linear combinations, 99_{n4}
- Lippman, Walter, 290
 on stereotypes, 15
- LISREL, 223, 245_{n2}
- Locksley, A., 294
 on individuating information, 372–74
- logical incoherence of defining stereotypes as inaccurate, 284
- low-certainty perceivers, 103–4
- low wattage view, 4–5
- LSATs, 216_{n8}
- MacKay, C., 192_{n5}
- Mackie, M., 313
 on stereotypes, 297
- Macrae, C. N., 373
 on person perception, 383
- Madon, S., 237–38, 240, 263
 on self-verification, 109
 on synergistic accumulation, 257–58
 on teacher expectations, 386
- Maehr, M., 54
- Major, B., 110_{n1}
 on self-verification, 105–7
- “Major Developments in Social Psychology during the Past Five Decades” (Jones), 148–49
- Malloy, T. E., 94
 social class stereotype study, 129–31
- Mannarelli, T., on person perception, 384–85, 387–88
- MANOVA. *See* Multivariate analysis of variance
- market predictions, 88, 192_{n5}
- Martel, R. F., 375, 376, 377
- Martinek, T., teacher expectation study, 224
- Marxism, 398
- math class, bias in, 227
- Mazella, R., sex stereotype meta-analysis, 137
- MBAs, accumulation of self-fulfilling prophecies and, 261–63
- MCATs, 216_{n8}

- McCain, John, 430
- McCauley, C., 169*n*7, 178, 225*t*, 393, 396–97
 racial stereotypes study, 324–27, 325*t*–326*t*
 sex stereotype study, 336*t*
- McFarlin, D. B., 105–7
- McGinnies, E., 21
- McMillan, D., 142*n*11, 142*n*12
 meta-analysis, 140
- McNulty, S. E., on self-verification, 107–8, 109
- Mead, Margaret, 313
- mean, regression to, 253
- Meehl, P. E., 181
- memory
 bias, 50*t*, 51*t*, 67–68, 85*t*
 Cohen study on, 67–68, 138–41
 expectations and, 122–42
 friendly, 71–72
 intelligent, 71–72
 labels and, 138
 Snyder and Uranowitz study on, 68, 139
 Stangor and McMillan meta-analysis, 140
 stereotypes and, 67–68, 138
- mental illness label
 effects of, 69, 136
 expectancy-confirming bias, 132–36
 Rosenhan on, 69–70, 132–35
- mentalistic concepts, 151
- Merton, R. K., 14, 61, 90
 on self-fulfilling prophecies, 26–28
- meta-analysis, 141*n*2
 defined, 359*n*3
 history of, 40
 on individuating information, 377
 Kunda and Thagard, 137, 378
 Mazella and Feingold, 137
 on stereotypes and person perception, 374–75
 Raudenbush, 44
 Rosenthal and Rubin, 40, 43
 of self-fulfilling prophecies, 408
 sex stereotypes, 137, 380
 Stangor and McMillan, 140
 Swim, 137, 359*n*3
 teacher expectations, 222–23, 408
- Mets, 124–25
- mice, 284
- Michigan Educational Assessment Program, 239
- Michigan Study of Adolescent Life Transitions (MSALT), 238
- Microsoft, 89
- Milgram, S., 185
- Military, U.S., 249
- military personnel, 99*n*1
- Milne, A. B., 373
- Milton, L. P., 108
- Miyake, K., 119–21
- moderators, 408–10
- Monte Carlo studies, 359*n*1
- moral beliefs, 9
- Morling, 350*t*, 397
 ballet dancer stereotype study, 352–54
- Morris, M. E., 384–85, 387–88
- MSALT. *See* Michigan Study of Adolescent Life Transitions
- multiculturalism, 288–89, 333–35, 399
- multiple-perceiver effects, 255
- Multivariate analysis of variance (MANOVA), 96
- Murphy, G., 20
- Murray, C., 37, 156
- mutual exclusivity, 7
- Myers, D. G., on stereotypes, 297, 298
- naive realism, 13
- NASDAQ, 88–90, 192*n*4
- National Basketball Association, 363
- National Election Study, 354
- national personality stereotypes, 399–400
 Heine study on, 399–400
 Terracciano study, 399–401
- Native Indians (Canadian), 331–33
- naturalistic studies, 40–41
 omitted variables in, 255
 on self-fulfilling prophecies, 254–55
 on teacher expectations, 223–24, 233
- The Nature of Prejudice* (Allport), 18
- near miss, 319
- Neuberg, S. L., 377, 427, 428
 on individuating information, 372–73
- New Jersey, 362
- New Look
 accuracy and, 22
 Allport, F. on, 22
 bias and, 22
 inaccuracy and, 22
 social perception and, 19–22
- New York, 364, 366
- New York Times*, 88

- 9/11, 9
- Nisbett, R. E., 11
- Nobel Prize, 413
- noncomponential approach, 204–5, 392
- nonconscious prejudices, 411
- nonindependent judges, agreement with, 185
- nonverbal behavior, gender stereotypes
 regarding, 339–40
- nonverbal sensitivity, 398
- Norton, M. I., 414, 415
- Oakes, J., 304
- Obama, Barack, 428, 432*n*2
 individuating information and, 430
 presidential election of, 430–32
 racial stereotypes and, 428, 430–32
- obedience, 185
- objective criteria, 18, 176–82
 consensual gender stereotypes with, 335–39
 imperfect, 181
 overall accuracy levels, 178
 standardized, 179–82
- objective stimuli, perception of, 14
- O'Brien, M., 241
- occupations
 sex stereotypes and, 342–43
 stereotypes, 348–54
- Olson, J. M., on MBAs and accumulation,
 261–62
- omitted variables, 255
- oppositional identity, 30
- oppression, 402
 Beyer on, 348
 overestimating, 348
- oppressive societies, 401
- optimism, 3
- Oracle, 89
- outgroup homogeneity, 357
- out-groups, 27–28
- overgeneralization, 357
- Pack, S. J., 56, 97
- Pakistani, 313, 331–33
- paranoid, 135
- Park, B., 204, 328, 333
 componential methodology, 392
 full accuracy design, 200–201
 on inaccuracy, 202, 317
 political stereotypes study, 354–55
 on relationship between stereotypes and
 prejudice, 298
 sex stereotype study, 349*t*
- pay, sex stereotypes and, 348–52
- peace activists, 365
- perceivers, 50*t*
 accuracy, 164–65, 198, see also teacher
 expectation accuracy
 agreement with, 184–87
 constrained, 113–14
 effect, 198
 expectations, 162, 253–61, see also
 expectancies, expectations,
 stereotypes, teacher expectations
 group effect, 200
 high-certainty, 104–5
 impressions, 165
 judgments, 146
 low-certainty, 103–4
 multiple-perceiver effects, 255
- perceptions. *See also* person perception
 African American students, of white
 students, 328–31
 expectancy effect sizes and, 85*t*
 expectations and, 122–42
 expectations bias person, 65–73, 125–38, see
 also stereotypes bias person perception
 interpersonal, 6–7
 of objective stimuli, 14
 of populations, 309–11
 stereotypes bias person, 65–70, 125–38,
 375–80
 white students, of African American
 students, 328–31
- perceptual defense, 19, 20–21
- peremptory challenge, 414–15
- peripheral traits, 25
- permissibility, 155
- personal constructs theory, 146
- personality traits, 388*n*2, 399–400
- personal stereotypes, 317, 321
 accuracy, 329–30, 325–26*t*, 329–30,
 331–33, 336–38*t*, 340, 345, 347,
 349–51*t*, 352–54, 355, 356
 accuracy of consensual stereotypes compared
 with accuracy of, 395–96
 correspondence with real differences,
 329–30, 333, 340–41, 345, 347,
 352–54, 355, 398

- discrepancies, 331–32, 347, 356
- gender, 336–38*t*, 340–41, 345, 347–48
- racial, 325–26*t*, 329–30, 331–333
- self-reports and, 352
- person perception
 - expectations bias, 65–70, 72–73
 - group differences in, 386
 - meta-analyses of stereotypes and, 374–75
 - stereotype accuracy and, 305*n*2, 311
 - stereotypes and, 357–59, 360–88, 391
 - stereotypes bias, 65–70
 - teacher expectations and, 386
- pervasive stereotype accuracy, 323–59
- pessimism, 3
- phenomenon, defining, 291
- physical attractiveness stereotype, 58–59, 95–97
 - accumulation and, 261–62
 - salaries and, 261–62
- Pinderhughes, E., 288
- polemics, 145
- polite/blunt, 25
- political beliefs, 9
 - accuracy and, 152–55
 - sex stereotypes and, 346–48
 - social sciences and, 156, 424
 - Stangor on, 154
 - stereotypes and, 296–97
- political stereotypes, 397
 - Judd and Park study, 354–55
- politicians, 367–68
- Polzer, J. T., 108
- Popper, K. R., 163, 296–97, 306*n*7
- populations, perceptions of, 309–11
- Portuguese, 331–33
- positive expectations, 37
- positive psychology, 280
- positive test strategy, 119
- prediction, 164–165
 - psychology of, 152, 366
- prejudice, 15–16, 18, 139, 250
 - belief similarity model of, 371
 - nonconscious, 411
 - stereotypes and, 299, 304, 391
- prescriptive beliefs, 9
- presidential elections, 177
 - accuracy and, 191*n*1
 - Obama in, 430–32
- pretest ratings, 31–32
- previous achievement, 239–40
- priming, stereotype, 410
- prisoner's dilemma, 61
- probabilistic realism, 171, 191
 - accuracy criteria, 171–73
 - social beliefs and, 172
 - social reality and, 172
- probabilistic stereotypes, 312
- process, 201–2
- “Process Affecting Scores on ‘Understanding of Others’ and ‘Assumed Similarity’” (Cronbach), 201
- processistic fallacy, 394–95
- prosecution, 413–14
- Prozac, 135
- psychiatric diagnoses, 69–70, 135
- psychological hypothesis, social
 - stereotypes compared to, 393*t*
- psychology
 - behaviorism, 148, 151
 - cognitive, 11–12, 12*t*, 151–52
 - cross-cultural, 151
 - cultural, 288–89
 - educational, 79*n*4
 - experimental, 33
 - positive, 280
 - of prediction, 366
 - social, 11–12, 12*t*, 51–52, 63*n*2, 76–77, 146, 149–50, 169*n*5, 273–74, 410
- Psychology Review*, 305*n*2
- Public Opinion* (Lippman), 15
- Pygmalion studies, 30–46, 79*n*4, 220–21, 237
 - acceptance of, 36–37
 - accumulation issues and, 259
 - boot camp, 61
 - caveats to, 44–45
 - controversy, 45
 - critical evaluation of, 92
 - criticism, 38–39
 - follow-ups, 39–40
 - interpreting, 39
 - IQ tests effects and, 33–39, 41–43, 45
 - meta-analysis, 40
 - naturalistic studies and, 40–41
 - reactions to, 36–39
 - re-analysis of, 43–44
 - of Rosenthal and Jacobson, 33–37, 39–40, 65, 220–21, 259, 424
 - teacher expectations and, 33–36

- Al Qaeda, 365
- “Quality of the Stimulus Tapes” (Duncan), 128
- questions, 112–121
- bias in, 74–76, 114
 - confirmatory, 115
 - constraining, 113–14
 - diagnostic, 117–18, 121, 424
- questionnaires, 190
- quiz show study, 78*n*1
- racial stereotypes, 53–54, 66–67
- accuracy of, 324–35, 325*t*, 326*t*
 - Ashton and Esses study, 325*t*–326*t*, 331–33
 - Chen and Bargh study, 410–12
 - consensual, 327
 - expectancy-confirming bias studies, 126–29
 - McCauley and Stitt study, 324–27, 325*t*–326*t*
 - Obama and, 428
 - peremptory challenge and, 414–15
 - personal, 329–30
 - Ryan study, 325*t*–326*t*, 328–31
 - self-fulfilling prophecies and, 54
 - unconscious, 429
 - Wolsko study, 333–35
- racism, 34
- RAM. *See* Realistic Accuracy Model
- random variation, 142*n*7
- random walk aspect, 186
- rational judgment, 416
- Raudenbush, S. W., 42–43, 43*f*
- meta-analysis of, 44
- realism
- defined, 171
 - naive, 13
 - probabilistic, 171–73, 191
- Realistic Accuracy Model (RAM), 210–13
- reasonableness, 6
- bias and, 416
 - data on, 370–88
- red herring arguments, 155
- reference group effect (RGE), 399
- regression, 223, 247*n*10
- coefficients, 247*n*8, 247*n*13
 - to mean, 253, 264*n*1
 - models, 387
- reign of error, 61
- rejection sensitivity, 412–13
- religion, 286
- Republicans, 354
- response bias, 195
- Rettew, D. C., 402
- RGE. *See* reference group effect
- Right Wing Authoritarianism (RWA) scale, 332, 341, 399
- Rist, R., 57, 62, 250
- critical evaluation of, 92
 - on social class stereotypes, 226, 259
- Rokeach, M., 372
- romantic relationships, 412–13
- Roselli, F., on stereotypes, 297
- Rosenberg, S., 146
- Rosenthal, D. L., 142*n*9, 142*n*10, 250, 423
- on mental illness label, 69–70, 132–35
- Rosenthal, R., 7, 45, 46, 51, 52, 62, 63*n*2, 190, 228, 254, 419
- binomial effect size display, 228, 320, 392–93
 - critical evaluation of, 92
 - on experimenter effects, 31–32
 - on IQ spurts, 226
 - meta-analysis, 40, 43
 - Pygmalion study of, 33–37, 39–40, 65, 220–21, 259, 424
- Ross, Lee D., 11, 75, 305*n*2
- on person perception, 384
 - quiz show study, 78*n*1
- Rothbart, M., 294
- friendly/intelligent memory study of, 71–72
- Roxbury Prep, 277–79, 305*n*4
- curriculum, 278
- Rozin, P., 393
- Rubin, D. B., 52
- meta-analysis, 40, 43
- Rubovitz, R., 54
- Ruderman, A., 294
- Russell, J., on MBAs and accumulation, 261–62
- RWA. *See* Right Wing Authoritarianism scale
- Ryan, C. S., 350*t*, 359*n*2
- on inaccuracy, 202
 - racial stereotype study, 326*t*–327*t*, 328–31
 - sorority study, 355–56
- Safin, Marat, 84
- Sagar, H. A., 128–29
- Sampras, Pete, 84
- Sarrazin, P., teacher expectation study, 224

- SATs, 45, 46*n*5, 180, 216*n*8, 239, 247*n*9, 249, 255, 374
- African American student scores, 419*n*1
- score differences, 419*n*1
- white student scores, 419*n*1
- schema concept, 151, 410
- self-schemas, 146
- schizophrenia, 132–33, 135
- Schmidt, F. L., 181
- Schneider, D. J., 304
- on social perception, 149–50
- on stereotypes, 299
- Schofield, 128–29
- the self, 10
- self-conception, 103
- self-esteem, 101–2
- self-fulfilling competition, 59–60
- self-fulfilling prophecies, 6–8, 10, 26–28, 120, 403, 421
- accumulation of, 248–65
- accumulation of, over time, 258–59
- accuracy and accumulation of, 252
- accuracy compared to, 160–65
- African Americans and, 26–27
- in agreement, 187
- Alvidrez and Weinstein study on, 243–44
- assessment of, 161
- Babad study on, 241
- bank insolvency and, 90–91
- Chen and Bargh study on, 99*n*3, 410–12
- in classroom, 222–24
- concurrent accumulation effects, 253–58, 256*f*, 263
- critical evaluation of, 91–97
- double standards, 166–67
- downtrodden, 237
- effect size and, 109
- expectancy effect sizes and, 85*t*
- experimenter effects and, 31–33
- group inequalities and, 421
- Hinnant, O'Brien, and Ghazarian study on, 241
- inadequacy of, 422–26
- information-seeking and, 75
- labor unions and, 26–27
- limited nature of, 83, 240–44
- logic of accumulating, 251–52
- Merton on, 26–28
- meta-analysis of, 85–86*t*, 41, 42–44, 408
- among military personnel, 99*n*1
- moderators, 408
- naturalistic studies, 254–55
- new phenomenon, 408–13
- perceivers in, 164–65
- potential limitations to accumulation of, 252–53
- potential mediators of, 50*t*
- power of, 49–63
- processes, 407–8
- racial stereotypes and, 54
- regression to mean and accumulation of, 253
- rejection sensitivity, 412–13
- scientific validity of, 163–64
- self-verification and, 100–111
- self-verification and accumulation of, 252
- social class and, 239
- social injustice and, 421
- social psychology and, 51–52
- in sports, 84–87
- stereotype priming and, 410–12
- stereotype threat and, 408–10
- stock market and, 87–89
- student stigma and, 238
- Sutherland and Goldschmid study on, 242–43
- synergistic accumulation and, 257–58
- teacher expectations and, 44, 407–8
- teachers' pets and, 219–20
- teacher-student relationships and, 235–36
- troublemakers and, 219–20
- self-fulfilling stereotypes, 52–59, 91–97
- "The Self-Fulfillment of the Self-Fulfilling Prophecy" (Wineburg), 42–43
- self-perception, 110*n*2, 187
- clarity of, 408
- componential models, 204
- expectations and, 408
- targets, 188–89
- self-reports, 330
- componential models, 204
- personal stereotypes and, 352
- scales, 190
- target, 187–88
- self-schemas, 146
- self-verification
- accumulation and, 252
- explained, 101–2
- influence of, 109

- self-verification (*Contd.*)
 Madon study on, 109
 Major, Cozzarelli, Testa, and McFarlin on, 105–7
 McNulty and Swann study on, 107–8, 109
 research on, 105–9
 self-fulfilling prophecies and, 100–111
 student, 234–35
 Swann, Milton, and Polizer study on, 108
 Swann and Ely on, 102–5, 109, 110, 120–21
- sensitivity, social, 341
- sexist interviewer study, 55–56
 critical analysis of, 97
- sex stereotypes, 54–59, 67, 316, 402
 academic achievement and, 343–44
 accuracy, 335–48
 Beyer study, 337*t*, 343–46
 Briton and Hall study, 336*t*, 339–40
 Cejka and Eagly study, 337*t*, 348–52, 349*t*
 in classroom, 56–57
 consensual, 335–39
 Diekman study, 337*t*, 346–48
 Eagly on, 336*t*, 346–48
 expectancy-confirming bias and, 131–32
 Goldberg study, 131–32
 Hall and Carter study, 337*t*
 individual differences in, 340–42
 Judd and Park study, 349*t*
 Kunda and Thagard analysis, 137
 Mazella and Feingold analysis, 137
 McCauley and Thangavelu study, 336*t*
 meta-analysis, 137, 378, 380
 nonverbal behavior and, 339–40
 occupations and, 342–43
 pay and, 348–52
 political beliefs and, 346–48
 Swim study, 336*t*
 teacher expectations and, 387
 wages and, 348–52
- sexual orientation, 418–19
 stereotype accuracy and, 419
- Sharpton, Al, 432*n*2
- Sherman, S. J., 79*n*5
 on confirmation, 118
 on diagnosis, 118
- simulation studies, 375
- skill, students', 234
- Skov, R. B., 79*n*5
 on confirmation, 118
 on diagnosis, 118
- Skrypnik, B. J., 54–55
- small bias accumulation, 375
- Smith, Alison, 237–38
- Snow, R. E., 38, 43–44
- Snyder, M., 54–55, 58–62, 112–13, 294, 304, 423–24
 on bias hypothesis, 118
 critical evaluation of, 95–97
 information-seeking studies, 113–14
 on memory, 139
 social hypothesis testing, 74–75
 stereotype studies, 68
- soccer coaching, 123–24
- social beliefs
 accuracy and, 169*n*8
 probabilistic nature of, 172
 social realities and, 11–12
- social class stereotypes, 57–58, 142*n*6, 180
 Baron, Albright, and Malloy study, 129–31
 in classroom, 259–60
 Darley and Gross study, 67, 129–31
 expectancy-confirming bias, 65–73, 129–31
 previous achievement and, 239–40
 Rist on, 57–58, 92–93, 226, 259
 self-fulfilling prophecies and, 57–58, 92–93, 226, 239, 259
- social condition, improving, 154
- social constructivism, 173
 social reality and, 174
- social hypothesis testing, 74–76, 112–121
 Snyder and Swann, 74–75, 112–114, 118
- social influence, 395, 410
- social information gathering, 118–21
- social injustice, 52
 expectancy and, 153
 self-fulfilling prophecies and, 421
 stereotypes and, 16–17
- social judgment, componential
 approaches to, 203*t*
- social nature, 5
- social problems
 alleviation of, 156–57, 244
 exacerbation of, 279
 self-fulfilling prophecies as sources of, 62, 422
- social psychology, 11–12, 146
 on accuracy research, 149

- bias and, 12*t*, 169*n*5
- expectancy-induced biases and, 76–77
- interpersonal expectations and, 63*n*2
- self-fulfilling prophecies and, 51–52
- trends in, 410
- social reality, 5–6, 13–29
 - interpersonal perceptions and, 6–7
 - judgment and, 14
 - perception and, 14
 - probabilistic nature of, 172
 - social constructivism and, 174
- social relations model (SRM), 197–200
- social sciences
 - data in, 426–28
 - on group differences, 391
 - on person perception, 368–70
 - politics in, 156, 424
 - on stereotype accuracy, 313
 - on stereotypes, 402
- social sensitivity, 341
- Sommers S. R., 414, 415
- sorority stereotypes, 355–56
- S&P 500, 90, 192*n*4
- sports, 99*n*2, 284–87
 - self-fulfilling prophecies in, 84–87
- Spoth, R., on synergistic accumulation, 257–58
- SRM. *See* social relations model
- Stahelski, A. J., 59
- standard deviations, 322*n*4, 344, 374
- standardized tests, 10, 216*n*8
- Stangor, C., 142*n*11, 142*n*12, 169*n*6
 - meta-analysis, 140
 - on political beliefs, 154
- statistical significance, 224
- Steele, C. M., 10, 238, 409, 419*n*1
- Steinmetz, J. L., quiz show study, 78*n*1
- stereotypes, 8–9. *See also* consensual stereotypes;
 - racial stereotypes; sex stereotypes
- absolutist, 312
- African Americans, 324–27, 411
- Allport, F., on, 285–86, 293, 303–4
- Allport, G., on, 18–19
- of Armenians, 16–18
- Ashmore and Del Boca on, 302
- ballet dancer, 352–54
- benefits of, 300–301
- bias and, 18–19, 290–91
- as biased expectations, 66
- business major, 353–54
- Claire and Fiske on, 253–54, 375–76
- classic media, 276*f*
- classic view of, 281–82, 290
- connotations of, 303–4
- counterstereotypic attributes, 330
- creation of, 273
- critiques of, 274–75
- defined, 271–72, 296–97, 301–13
- discrimination and, 299, 391
- distinguishing, 389–90
- dynamics of, 390
- early work on, 15–19
- eliminating, 380
- engineering major, 353–54
- entrenched, 412
- error in, 290–91
- evidence for, 307–22
- evil and, 279–80
- exclusions from, 302–3
- expectations based on, 52, 125–41
- Fiske and Taylor on, 297, 298
- full accuracy design, 200–201
- group difference acceptance, 284–87, 312–16
- harmful effects of, 281–82
- hypocrisy and, 282–89
- inaccuracy and, 16, 269–306, 391–92, 397–402
- individuating information and, 378–79
- inequalities and, 280
- irrationality and, 16, 375–77
- judgment and, 150, 370–88
- Katz and Braly on, 15–16, 313, 315, 317–18
- LaPiere on, 16–18
- Lippman on, 15, 290
- logical incoherence and, 284
- Mackie, Hamilton, Susskind, and Roselli on, 297
- memory and, 67–68, 138
- mice and, 284
- models, 193*n*6
- Myers on, 297, 298
- national personality, 399–401
- negative effects of, 279–81, 299, 376
- neutral definitions of, 297, 302
- occupational, 348–54
- oppression and, 402
- in oppressive societies, 401
- origins of, 391
- as perceptions of populations, 309–11

stereotypes, (*Contd.*)

- personal, 317, 321, 329–33, 340, 345, 347, 352, 395–96
 - person perception and, 357–59, 360–88, 391
 - person perception biased by, 65–70
 - physical attractiveness, 58–59, 95–97
 - political, 354–55, 397
 - political beliefs and, 296–97
 - powerful effects, 375–77
 - prejudice and, 299, 391
 - priming, 410–12
 - probabilistic, 312
 - processistic fallacy, 394–95
 - psychological benefits of, 296–97
 - reliance on, 373, 378, 385, 391
 - Roxbury Prep and, 277–79
 - self-fulfilling, 52–53
 - shared components of, 396
 - Snyder on, 68
 - social, 251, 393*t*
 - social class, 57–58, 142*n*6, 180
 - social injustice and, 16–17
 - in social psychology, 273–74
 - social science perspectives on, 402
 - sorority, 355–56
 - of stereotypes, 402
 - straw arguments, 282–83
 - target confirmation, 409
 - threat, 10–11, 179, 308, 408–10, 420*n*1
 - Uranowitz on, 68
 - use of, as term, 303–4
- stereotype accuracy, 155, 158–59, 196, 273–74, 279–81, 305*n*2, 321*n*3, 393
- aspects of, 316–18
 - bull's eye, 318–19
 - caveats, 321
 - clarifications, 321
 - corresponding with differences, 317, 320–21
 - discrepancies, 318–20
 - discrepancy from perfection in, 316–17
 - discrepancy types, 319–20
 - early hints of, 314–15
 - evidence of, 307–22
 - exaggeration hypothesis and, 314
 - high, 320–21
 - in individual judgment, 361
 - levels of analysis, 308–12
 - limitations of, 396–97
 - near miss, 319

- personal, 329–30
 - person perception and, 311
 - racial, 324–35, 325*t*, 326*t*
 - results of, research, 390–97
 - scientific value of, research, 389–90
 - sexual orientation and, 419
 - in small groups, 311
 - social science scholarship on, 313
 - social value of, research, 389–90
 - standards for, 318–21
 - teacher expectations and, 386–88
- Stereotype Rationality Hypothesis, 380–81, 385–86, 388
- stigma, 238
- Stinson, L., 418
- Stitt, C. L., 178
- racial stereotypes study, 324–27, 325*t*–326*t*
- stock market, 177
- self-fulfilling prophecies and, 87–89
- straw arguments, 282–83
- students. *See also* teacher expectations
- African American, 328–31, 419*n*1
 - competence and, 234
 - intelligence and, 234
 - personal situations and, 237
 - self-verification and, 234–35
 - skill and, 234
 - stigmas, 238
 - teacher expectations and, 231, 232*t*
 - teacher-student relationships, 235–36
 - white, 328–31, 419*n*1
- Surowiecki, J., 395, 396
- Susskind, J., on stereotypes, 297
- Sutherland, A., on self-fulfilling prophecies, 242–43
- Swann, W. B., Jr., 60–61, 110*n*2, 111*n*4, 112–15, 423–24
- on bias hypothesis, 118
 - information-seeking studies, 113–15
 - predictions from, 103*t*
 - on self-verification, 102–5, 107–10, 120–21
 - social hypothesis testing, 74–75
- Swim, J. K., 131*f*
- meta-analysis, 137, 359*n*3
 - sex stereotype study, 335–39, 336*t*
- synergistic accumulation, 257–58
- system justification, 12*t*, 393

- taboo words, 20–21
- tachistoscopes, 20–21
- Tanke, E. D., 58, 294
- critical evaluation of, 95–97
- targets, 122
- agreement with, 187–89
 - group effect, 200
 - self-perception, 188–89
 - self-reports, 187–88
 - stereotype confirmation and, 409
- Taylor, S. E., 54, 294
- on stereotypes, 297, 298
- teacher expectations, 219–47, 421.
- See also* students
 - accuracy, 230–34
 - assessing accuracy of, 231–34
 - Brattesani study on, 223–24
 - causal ambiguity of, 231
 - effect size, 222–23, 225*t*
 - experimental studies on, 222
 - inaccurate but uninfluential, 231
 - Jussim and Eccles study on, 227
 - Jussim studies on, 224, 227, 386
 - limitations of, 227–30
 - Madon study on, 386
 - meta-analysis, 222–23, 408
 - naturalistic studies on, 223–24, 233
 - person perception and, 386
 - power of, 221–22
 - processes, 407–8
 - Pygmalion studies and, 33–36
 - race and, 238–39, 386–87
 - self-fulfilling effects of, 33–36, 222–30, 237–44, 259–61, 407–8
 - sex stereotyping and, 386–87
 - social class and, 57–58, 92–93, 226, 237, 239–40, 386–87
 - stereotype accuracy and, 386–88
 - student achievement and, 231, 232*t*
 - teachers' own, 230–34
 - teachers' pets, 219–20
 - troublemakers, 219–20
 - Trouilloud, Sarrazin, Martinek, and Guillet, 224
 - weakness of, 230
 - West and Anderson study on, 223
 - Williams' study on, 72, 226–27
- teachers' pets, 219–20
- teacher-student relationships, 235–36
- Terracciano, A.,
- national personality stereotype study, 399–401
- Testa, M., 105–7
- Thagard, P., sex stereotype meta-analysis, 137, 378, 380
- Thangavelu, K., 393
- sex stereotype study, 336*t*
- thick slices, 419
- thin slices, 418–19
- TOGA. *See* Flanagan's Test of General Ability
- "Topic Areas Often Covered by Interviewers" (Snyder & Swann), 74
- treatment blind, 128
- Trope, Y.
- on confirmation, 117
 - on diagnosis, 117
 - on hypothesis-testing, 117
 - information-seeking study, 114
- troublemakers, 219–20
- Trouilloud, D., 227
- teacher expectation study, 224
- truth, accuracy criteria and, 175–76
- Tversky, A., 366, 370
- unbiased information gathering, 73–74
- unbiased judgment, 416
- unconscious racism, 429
- Understanding Race, Ethnicity, and Power* (Pinderhughes), 288
- United States, 401
- universalism, 341, 399
- University of Colorado, 328, 333–35
- Uranowitz, S. W.
- on memory, 139
 - stereotype studies, 68
- urban schools, bias in, 226
- valence, 417
- accuracy, 418
 - patterns for, 418
- variables, omitted, 255
- Vietnamese, 331–33
- von Hippel, Bill, 281
- wages, sex stereotypes and, 348–52
- Wall Street Journal*, 88
- warm/cold, 25
- Weather Channel, 170, 362

- Webb, Spud, 363
- Weinstein, R. S., 225*t*
 on self-fulfilling prophecies, 243–44
- West, C., 260
 on teacher expectations, 223
- “Who Wants to Be a Millionaire?” (television), 395
- Wilder, D. A., 304
- Willard, J., 257–58
- Williams, T., 93, 225*t*, 227
 teacher expectation study, 72, 226–27
- Wilshire 5000, 90
- Wineburg, S. S., on IQ tests, 42–43
- The Wisdom of Crowds* (Surowiecki), 395
- Woll, S. B., 139
- Wolsko, C., 399
 racial stereotypes study, 333–35
- Word, C. O., 53, 62, 94, 99*n*3
 critical evaluation of, 93–95
- World Series, 124–25
- Yankees, 124–25
- Zanna, M. P., 53, 55, 97
 critical evaluation of, 93–95
- zero acquaintance, 418–19
- Z scores, 40, 46*n*2
- Zuckerman, M., on bias hypothesis, 119–21